DOCTORAL THESIS

Testing spoken language using computer technology

A comparative validation study of 'live' and computer delivered test versions using Weir's framework

Zainal abidin, Saidatul Akmar

Award date:
2006

TESTING SPOKEN LANGUAGE USING COMPUTER TECHNOLOGY: A COMPARATIVE VALIDATION STUDY OF 'LIVE' AND COMPUTER DELIVERED TEST VERSIONS USING WEIR'S FRAMEWORK

by
Saidatul Akmar Zainal abidin

A thesis submitted in partial fulfilment of the requirements for the degree of PhD

Centre for Language Assessment Research

School of Arts

Roehampton University

University of Surrey

2006

ABSTRACT

The purpose of this research is to investigate the validity of testing spoken language using computer technology; an area, which has to date, received very little attention in the testing literature. The focus on this new channel for testing speaking arose from the pressing needs of Mara University of Technology, Malaysia, where over 10,000 full-time and part-time students located at various locations across the country take the test every year. This direct method of delivering a spoken test to these students presents the university with serious problems in terms of efficiency, reliability and validity. These are significant aspects for any test, and to date, they have not been addressed formally at the university.

To achieve this, a new framework for validating a speaking test (Weir 2004/2005) was adopted throughout the study. The framework was operationalised such that data collection and analysis were conducted according to validity elements of the framework, and consequently all findings were systematically reported. The study involved three phases: a validation study on the direct test, the development, administration and validation of a computer test, and a comparative analysis of the two methods of testing speaking.

Data from the first validation study point to problems faced by the direct test such as equivalent forms, the rating process, co-construction of discourse and rating in the interactive task, and other administrative & security issues. With the computer test, these problems are addressed, and since it paralleled the direct test in its contextual features, processing and performance were not affected in a drastic manner. A comparison of the two methods for testing speaking enables us to address the question of whether the computer test is a valid replacement of the direct test.

In so far as appropriate measures are taken from its development stage through to its administration, the computer speaking test is a viable alternative. Without proper steps to ensure its validity 'a priori', during and 'a posteriori' of the test event, all effort in developing the test will fail.

ACKNOWLEDGEMENTS

(In the name of Allah most Gracious most Merciful)

Now is the time to thank those who had provided me the opportunity to do this PhD, supported me and believed in me.

First, I would like to thank the Department of Civil Service Malaysia, who awarded me the scholarship, and my employer Mara University of Technology Malaysia, who approved a four year leave for completion of the PhD.

My biggest and deepest appreciation goes to my supervisors.
Professor C J Weir had inspired me and provided guidance for the project even before I came to the U.K. He gave me help and advice from the start, and constantly throughout the first two years in Roehampton. Thank you for your patience and support.

Dr Barry O'Sullivan, who took over this past year, had been there when I needed him. He read my chapters patiently and provided sound advice and criticisms. The completion of the thesis would not have been possible without his care and attention.

Thank you too to Dr Stephen Bax who read the thesis within a limited time at the end, and provided some constructive and most helpful suggestions.

I'd like to express my sincere thanks to all those who took part in my twice-a-year three years data collection process in Malaysia, especially to my colleagues and students. Those trips to Malaysia had been worthwhile especially because it gave me the chance to interact with students, who participated in the computer trials and tests with great enthusiasm and without question.

My most heartfelt gratitude goes to my beloved family who had given me unconditional support throughout this undertaking. This especially goes to my son Anis, who had

demonstrated enormous patience and courage in this longest number of years that we've been separated.

A final note goes to the late Dr Zaleha Salleh; a true friend, confidante and someone I turned to in times of solace. Though she is no more, I dedicate this work to her for I know she is proud of this success.

This endeavour has given me vast knowledge and the inspiration to continue working, but most of all, it has added a special meaning to my life, and for this I thank you all.

TABLE OF CONTENT

PAGE

ABSTRACT i
ACKNOWLEDGEMENTS iii
TABLE OF CONTENTS v
LIST OF FIGURES xii
LIST OF TABLES xiv
LIST OF ABBREVIATIONS xix

Chapter 1: INTRODUCTION

1.1 The Research Proposal 1
An overview of language programmes & language testing at the university:

A. The Malaysian context 2
B. The MUET exam 3

1.2 Context of the study 5
Language testing at Mara University of Technology:

A. Testing of Speaking at the university 5
B. Rationale of the study 6

1.3 Organization of the thesis 9

Chapter 2: REVIEW OF THE LITERATURE

2.1 Introduction 14

2.2 The spoken discourse 17

2.2.1 Theories on spoken language 19

2.2.1.1 In the field of Psycholinguistics 19

2.2.1.2 In the field of Linguistics and Applied Linguistics... 22
2.2.1.3 Other works on spoken discourse..... 24

2.3 Testing the spoken language 30

2.3.1 The direct speaking test 30

2.3.2 The semi-direct speaking test 40

2.3.3 Computerized testing 45

2.3.3.1 Computer-based tests 46
2.3.3.2 Computer adaptive tests..... 47
2.3.3.3 Tests on the web 48

2.4 The Socio-cognitive framework for validating skills in language tests 51

2.4.1 A model/framework for validating a speaking test.. 51

2.42 Components of the framework for validating the speaking test 55

a) Theory-based validity 56
b) Context validity 58
c) Scoring validity..... 59
d) Consequential validity 59
e) Criterion-related validity..... 60

2.5 Issues of Validity 61

A. Defining validity..... 61
B. Types of validity..... 63
C. Validating the construct 65

2.6 Research questions 67

Chapter 3: METHODOLOGY & DESIGN

3.1 Introduction 70

Section I: Research design according to research questions 75

Section II: Development of research instruments.. 85

3.2 Instruments for Main Study 1 86

A. The Speaking test 86

1. The Speaking test Questionnaire..... 86

a) Pilot study 1..... 90
b) Pilot study 2..... 96

2. The Speaking test Interview..... 105

a) Interview Trials..... 108

b) Developing 'Guidelines' and questions for the Interview 116

3. The Test Documents 117

4. Observation 120

3.3 Instruments for Main Study 2 121

A. Pre-trials of the monologue test..... 122

1. The Monologue Test 124

2. The Semi-direct 'Interactive' task..... 125

a. Preliminary design and trials 125
b. The Speaking test script and design 126

c. Criteria for rating the Speaking Test 131

d. Questionnaire for the CB Speaking Test... 133

B. Pilot study of computer- delivered test (both tasks A & B) 134

1. The computer test administrations..... 136
2. Presentation on developing rating criteria... 138
3. Conclusion from pre-trials and pilot study.. 139
4. Staff workshop on the testing of speaking... 140

Section III: Description of participants, locations and research schedule..... 149

Section 1V: Model of the research plan..... 153

Chapter 4: THE DIRECT TEST VALIDATION

4.1 Introduction..... 155

4.2 The Speaking Test..... 158

4.3 Findings Main Study 1..... 159

4.3.1 Context Validity..... 162

4.3.2 Overall comment on task settings..... 191

4.3.3 Overall comment on task demands 215

4.3.4 Conclusions from context validity analysis..... 216

4.3.5 Theory-Based Validity..... 218

4.3.6 Overall comment on executive processes..... 229

4.3.7 Overall comment on executive resources..... 247

4.3.8 Conclusions from theory-based validity analysis 249

4.3.9 Scoring Validity 250

4.4 Overall comment on scoring validity..... 266

4.5 Summary of Validity of the Direct Speaking Test 269

Chapter 5: THE SEMI-DIRECT COMPUTER TEST

5.1 Introduction 272

5.2 Underlying Principles 275

5.2.1 Limitations of the direct speaking test..... 277

5.3 Model of the test Design and Operations..... 279

5.3.1 Test design phase..... 279

A. The monologue test 279
B. The interaction test 280
C. Turns, length and functions..... 283
D. The computerized speaking test..... 283

5.3.2 Test operations phase..... 284

A. Pre-trials 285
B. Pilot Study 286
C. Main Study 2..... 288

5.4 Findings from Pilot Study and Main Study 2.... 290

5.4.1 Data from Pilot study..... 291

5.4.1.1 Context validity..... 292

5.4.1.2 Theory-based validity..... 293

5.4.2 Data from Main study 2..... 295

5.4.3 Context validity..... 297

5.4.4 Overall comment on task settings..... 313

5.4.5 Overall comment on task demands..... 332

5.4.6 Theory-based validity..... 334

5.4.7 Conclusion for executive process..... 346

5.4.8 Conclusion for executive resource..... 353

5.4.9 Scoring validity..... 355

5.4.10 Conclusion for scoring validity..... 358

5.5 Summary of the Validity of the Computer Test 359

Chapter 6: DISCUSSIONS: A Comparison of Findings from the Direct Test & the Computer Test Studies

6.1 Introduction 363

6.2 Comparison of Findings According to the Validity Components 366

A. Context validity 367

B. Theory-based validity 371

C. Scoring validity 378

Chapter 7: CONCLUSIONS

7.1 Introduction 382

7.2 Research Question 1 385

7.2.1 Discussion 386

7.3 Research Question 2 394

7.3.1 Discussion 396

7.4 Final Question/ Discussion 404

7.5 Implications of the Study 405

7.6 Limitations of the Study 408

7.7 Contributions to Literature 409

7.8 Future Research 410

7.9 Concluding Remarks 413

BIBLIOGRAPHY 416

APPENDICES IN THESIS

Appendix 3.1: Questionnaire difficulty & comments
Appendix 3.2: Student questionnaire (direct test)
Appendix 3.3: Staff questionnaire (direct test)
Appendix 3.4: Student Interview Notes
Appendix 3.5: Staff Interview Notes
Appendix 3.6: Guidelines for the Interview
Appendix 3.7: Speaking test documents
- a) Course syllabus
- b) Test specifications
- c) Sample question paper
- d) Score sheets
- e) Instructions for the speaking test and assessment procedure
- f) Criteria/Rating scale

Appendix 3.8: Computer test script Task B
Appendix 3.9: Computer test questionnaire: CV and TBV
Appendix 3.10: Report workshop May2005
Appendix 3.11: Computer speaking test script A & B
Appendix 3.12: Computer test specifications

APPENDICES IN CD I

Appendix 3A: Validity evidence table
Appendix 3B: Pilot data March 2003 & pilot data September 2003
Appendix 3C: Extract from March 2003 pilot data SPSS output
Appendix 3D: The monologue test
Appendix 3E: Developing criteria for rating a speaking test
Appendix 3F: Staff questionnaire SPSS data (May 2005 workshop)
Appendix 3G: Rating the direct test (using old scale and TOEFL)
Appendix 3H: Main Study 2 findings (Pilot + MS2)
Appendix 3I: Student computer familiarity data
Appendix 3J: Rating the direct test + computer test (September 2005)
Appendix 3K: IRR direct test and computer test (September 2005)
Appendix 3L: Student questionnaire data on computer test & direct test (SPSS)
Appendix 3M: Main Study 1 findings summary
Appendix 3N: Transcripts student and staff interview Main study 1

APPENDICES IN CD II: The computerized speaking test

1. in Power Point
2. in WAVE Audio files (Task A + Task B)

LIST OF FIGURES

Chapter 2

Figure 2.1 A blueprint for the speaker (Levelt 1989) 21
Figure 2.2 A Socio-cognitive framework for validating speaking tests (Weir 2005) 54
Figure 2.3 Summary of alternative framework suggested by O'Sullivan (2000a) 55
Figure 2.4 Characteristics of the test taker (O'Sullivan 2000a) 56

Chapter 3

Figure 3.1 Summary Matrix: methods & research questions 82
Figure 3.2 Details of data gathering for Main Study 1 84
Figure 3.3 Details of data gathering for Main Study 2 85
Figure 3.4a Reliability analysis of questionnaire items 94
Figure 3.4b Factor analysis of questionnaire items 95
Figure 3.5a Factor analysis on questionnaire items Section A 101
Figure 3.5b Factor analysis on questionnaire items Section B 102
Figure 3.5c Factor analysis on questionnaire items Section C 102
Figure 3.6 Comparison between interview items and questionnaire items 115
Figure 3.7 Comparison between rating criteria in test scales 131
Figure 3.8 Concerns with the computerized speaking test 139
Figure 3.9 Operational Model of the Research Design 154

Chapter 4

Figure 4.1 Design of the study 156
Figure 4.2 Features of the direct speaking test 159
Figure 4.3 Socio-cognitive framework for validating a speaking test (Weir 2005) 161
Figure 4.4 Aspects of Context Validity for Speaking (Weir 2004) 163
Figure 4.5 Aspects of Theory-based Validity for Speaking (Weir 2004) 219
Figure 4.6 Aspects of Scoring Validity for Speaking (Weir 2004) 251
Figure 4.7 Model of the test process 268
Figure 4.8 Summary of validity of the direct speaking test 270

Chapter 5

Figure 5.1 Figure 5.1 Features of the computerized speaking test.....284
Figure 5.2 Model of computer test design and operations.....290
Figure 5.3 Aspects of Context Validity for Speaking (Weir 2004).....297

Figure 5.4 Aspects of Theory-based Validity for Speaking (Weir 2004).....334
Figure 5.5 Aspects of Scoring Validity for Speaking (Weir 2004).....354
Figure 5.6 Summary of validity of the computer test.....361

Chapter 6

Figure 6.1 A comparison of findings for the direct test (from Main Study 1) and the computer test (from Main Study 2) 366

Chapter 7

Figure 7.1 Data gathered for Main Study 1 385
Figure 7.2 Summary of findings for Main Study 1.....386
Figure 7.3 Data gathered for Main Study 2 395
Figure 7.4 Summary of findings for Main Study 2 396

LIST OF TABLES

CHAPTER 4

Context Validity
Elements of task setting:

Table 4.1a Purpose (participant) 164
Table 4.1b Purpose (document) 165
Table 4.1c Purpose (observation) 166
Table 4.2a Response format (participant) 167
Table 4.2b Response format (document) 169
Table 4.2c Response format (observation) 169
Table 4.3a Weighting (participant) 171
Table 4.3b Weighting (document) 172
Table 4.3c Weighting (observation) 172
Table 4.4a Known criteria (participant) 173
Table 4.4b Known criteria (document) 174
Table 4.4c Known criteria (observation) 175
Table 4.5a Order of items (participant) 176
Table 4.5b Order of items (document) 177
Table 4.5c Order of items (observation) 178
Table 4.6a Time constraint (participant) 179
Table 4.6b Time constraint (document) 180
Table 4.6c Time constraint (observation) 181

Elements of test administration:

Table 4.7a Physical conditions (participant) 183
Table 4.7b Physical conditions (document) 184
Table 4.7c Physical conditions (observation) 185
Table 4.8a Uniformity of administration (participant) 186
Table 4.8b Uniformity of administration (document) 187
Table 4.8c Uniformity of administration (observation) 187
Table 4.9a Security (participant) 189
Table 4.9b Security (document) 190
Table 4.9c Security (observation) 190

Elements of task demands:

Table 4.10a Discourse mode (participant) 194
Table 4.10b Discourse mode (document) 195

Table 4.10c Discourse mode (observation) 196
Table 4.11a Channel of communication (participant)....197
Table 4.11b Channel of communication (document) 198
Table 4.11c Channel of communication (observation) 198
Table 4.12a Length (participant)....199
Table 4.12b Length (document) 200
Table 4.12c Length (observation) 201
Table 4.13a Nature of information (participant) 202
Table 4.13b Nature of information (document) 203
Table 4.13c Nature of information (observation) 204
Table 4.14a Content knowledge required (participant) 205
Table 4.14b Content knowledge required (document) 206
Table 4.14c Content knowledge required (observation) 207
Table 4.15a Linguistic variables (participant).... 208
Table 4.15b Linguistic variables (document) 210
Table 4.15c Linguistic variables (observation) 211
Table 4.16a Interlocutor variables (participant) 212
Table 4.16b Interlocutor variables (document) 213
Table 4.16c Interlocutor variables (observation) 214

Theory-based Validity

Elements of executive process:

Table 4.17a Conceptualiser/Preverbal message (participant) 221
Table 4.17b Conceptualiser/Preverbal message (document) 223
Table 4.17c Conceptualiser/Preverbal message (observation)....223
Table 4.18a Linguistic formulator (participant)....224
Table 4.18b Linguistic formulator (document) 226
Table 4.18c Linguistic formulator (observation) 226
Table 4.19a Overt speech (participant)....227
Table 4.19b Overt speech (document)....228
Table 4.19c Overt speech (observation)....229

Elements of executive resource:

Table 4.20a Content knowledge (participant)....232
Table 4.20b Content knowledge (document) 233
Table 4.20c Content knowledge (observation) 234
Table 4.21a Grammatical knowledge (participant) 236
Table 4.21b Grammatical knowledge (document) 237
Table 4.21c Grammatical knowledge (observation) 238

Table 4.22a Discoursal knowledge (participant) 239
Table 4.22b Discoursal knowledge (document) 241
Table 4.22c Discoursal knowledge (observation) 241
Table 4.23a Functional knowledge (participant) 242
Table 4.23b Functional knowledge (document) 243
Table 4.23c Functional knowledge (observation) 244
Table 4.24a Sociolinguistic knowledge (participant) 245
Table 4.24b Sociolinguistic knowledge (document) 246
Table 4.24c Sociolinguistic knowledge (observation) 247

Scoring validity

Table 4.25a Criteria/Rating scale (participant) 252
Table 4.25b Criteria/Rating scale (document) 253
Table 4.25c Criteria/Rating scale (observation) 254
Table 4.26a Rater training (participant) 255
Table 4.26b Rater training (document) 256
Table 4.26c Rater training (observation) 257
Table 4.27a Standardization (participant) 258
Table 4.27b Standardization (document) 259
Table 4.28a Rating conditions (participant) 260
Table 4.28b Rating conditions (observation) 261
Table 4.29a Moderation (participant) 262
Table 4.30a Statistical analyses (participant) 263
Table 4.30b Statistical analyses (document) 264
Table 4.31a Grading & awarding (participant) 265
Table 4.31b Grading & awarding (document) 266

CHAPTER 5

Context Validity

Elements of task setting:

Table 5.1a Purpose (participant)....298
Table 5.1b Purpose (document)....298
Table 5.1c Purpose (observation) 299
Table 5.2a Response format (participant)....300
Table 5.2b Response format (document) 301
Table 5.2c Response format (observation) 301
Table 5.3a Weighting (participant)....302
Table 5.3b Weighting (document) 302
Table 5.3c Weighting (observation) 303

Table 5.4a Known criteria (participant) 303
Table 5.4b Known criteria (document) 304
Table 5.4c Known criteria (observation) 305
Table 5.5a Order of items (participant) 305
Table 5.5b Order of items (document) 306
Table 5.5c Order of items (observation) 307
Table 5.6a Time constraint (participant) 307
Table 5.6b Time constraint (document) 308
Table 5.6c Time constraint (observation) 308

Elements of test administration:

Table 5.7a Physical conditions (participant) 310
Table 5.7b Physical conditions (document) 310
Table 5.7c Physical conditions (observation) 311
Table 5.8a Security (participant) 312
Table 5.8b Security (document) 312
Table 5.8c Security (observation) 313

Elements of task demands:

Table 5.9a Channel of communication (participant) 316
Table 5.9b Channel of communication (document) 316
Table 5.9c Channel of communication (observation) 317
Table 5.10a The computer-delivered test (participant) 318
Table 5.10b The computer-delivered test (document) 319
Table 5.10c The computer-delivered test (observation) 319
Table 5.11a Nature of information in tasks A & B (participant) 320
Table 5.11b Nature of information in tasks A & B (document) 321
Table 5.11c Nature of information in tasks A & B (observation) 321
Table 5.12a Topic familiarity (participant) 322
Table 5.12b Topic familiarity (document) 323
Table 5.12c Topic familiarity (observation) 323
Table 5.13a Linguistic variables (participant).... 324
Table 5.13b Linguistic variables (document) 325
Table 5.13c Linguistic variables (observation) 325
Table 5.14a Interlocutor variables task A (participant) 327
Table 5.14b Interlocutor variables (document) 329
Table 5.14c Interlocutor variables (observation) 329
Table 5.15a Interlocutor variables task B (participant) 330
Table 5.15b Interlocutor variables (document) 331
Table 5.15c Interlocutor variables (observation) 332

Theory-based Validity
Elements of executive process:

Task A
Table 5.16a Conceptualiser/Preverbal message (participant) 335
Table 5.16b Conceptualiser/Preverbal message (observation).....335
Table 5.17a Linguistic formulator (participant).....336
Table 5.17b Linguistic formulator (observation) 337
Table 5.18a Planning time (participant).....337
Table 5.18b Planning time (observation).....338
Table 5.19a While speaking (participant).....339
Table 5.19b While speaking (observation).....340
Task B
Table 5.20a Conceptualiser/Preverbal message (participant) 340
Table 5.20b Conceptualiser/Preverbal message (observation).....341
Table 5.21a Linguistic formulator (participant).....342
Table 5.21b Linguistic formulator (observation) 342
Table 5.22a Planning time (participant).....343
Table 5.22b Planning time (observation).....344
Table 5.23a While speaking (participant).....345
Table 5.23b While speaking (observation).....346

Elements of executive resource:

Table 5.24a Internal (background) knowledge (participant) 348
Table 5.24b Internal (background) knowledge (observation) 349
Table 5.25a External knowledge (participant).....350
Table 5.25b External knowledge (observation) 350
Table 5.26a Functional knowledge (participant) 351
Table 5.26b Functional knowledge (observation) 352
Table 5.27a Sociolinguistic knowledge (participant) 352
Table 5.27b Sociolinguistic knowledge (observation) 353

Scoring validity

Table 5.28a Criteria/Rating scale (participant) 355
Table 5.28b Criteria/Rating scale (document) 356
Table 5.28c Criteria/Rating scale (observation) 357
Table 5.29 Rating Decisions (inter-rater agreement) 357
Table 5.30 Consistency (intra-reliability) 357

LIST OF ABBREVIATIONS

UiTM	Mara University of Technology
EDC	Educational Development Centre
MUET	Malaysian University English Test
ELS	English Language Services
FSI	Foreign Services Institute
CPE	Certificate of Proficiency in English
EFL	English as a Foreign Language
ACTFL	American Council on the Teaching of Foreign Languages
ILR	Interagency Language Roundtable
ASTP	Army Specialized Training Programme
CEEB	College Entrance Examination Board
MLA	Modern Language Association
TSE	Test of Spoken English
TEEP	Test of English for Educational Purposes
CAL	Centre for Applied Linguistics
SPEAK	Speaking Proficiency English Assessment Test
CBT	Computer-based Test
CBLT	Computer-based Language Test
CAT	Computer Adaptive Test
GRE	Graduate Record Examination
SAT	Scholastic Aptitude Test
OPI	Oral Proficiency Interview
SOPI	Simulated Oral Proficiency Interview
COPI	Computerized Oral Proficiency Instrument
WBT	Web Based Test
WAT	Web Adaptive Test
TOEFL	Test of English as a Foreign Language
APA	American Psychological Association
AERA	American Educational Research Association
NCME	National Council on Measurement in Education
CEFR	Common European Framework of Reference
IELTS	International English Language Testing Service
CELS	Certificate in English Language Skills
TAST	TOEFL Academic Speaking Test

TESTING SPOKEN LANGUAGE USING COMPUTER TECHNOLOGY: A COMPARATIVE VALIDATION STUDY OF 'LIVE' AND COMPUTER DELIVERED TEST VERSIONS USING WEIR'S FRAMEWORK

Chapter 1 INTRODUCTION

1.1 The Research Proposal

This research is on the testing of spoken language in English using computer technology. The purpose of the research is twofold. It will look at the application of a framework for test validation (Weir 2005), and to use this framework for making validation comparisons of a test delivered 'live' and over a computer.

The immediate focus on this new channel for testing speaking arises from the pressing needs of various centres at Mara University of Technology Malaysia: the Language Centre which develops the English programmes that all candidates enrol for; the distance-learning centre (EDC- Educational Development Centre) where students attend the university on a part-time basis via the distance learning mode; and the Off-Campus Programme where candidates are in employment and attend the university on a part-time basis; though the study will have implications for testing contexts beyond the one described here. All these centres offer English programmes to the candidates, and whether students are enrolled on a full-time or part-time basis, English is a compulsory subject. The English programmes also include 'speaking' as a major component and this is tested in the final examinations.

The speaking test is currently administered as a direct, face-to-face test to over 10,000 full-time and part-time students at various locations (branch campuses) across the

country every year. This direct method of delivering a spoken test, especially to part-time and distance-learning candidates, presents the university with serious problems in terms of efficiency, reliability and validity. This study hopes to address these concerns of spoken language testing which is currently conducted at the university using the direct method.

For the above reasons, this research will look first to the existing speaking test at UiTM and investigate various aspects of the test via a validation process using the Socio-cognitive Framework for Validating a Speaking test (Weir 2004/2005), and investigate further how best to improve on conditions which could in turn heighten its merit as a test that reflects a candidate's speaking ability.

A. The Malaysian context

In the Malaysian context, very little work has been done on the validation of direct or indirect tests of speaking, with almost no research conducted on computer-based testing of any language skill and none on the testing of speaking. At present, there are two universities that offer a web-based learning mode: Mara University of Technology Malaysia/UiTM (the university where the researcher is based) and The Open University Malaysia. While speaking is found in the syllabuses, no evidence of validation or in-depth research work has been recorded at both universities on how language is tested, especially on testing of speaking, either using the direct or indirect method.

The same pattern where 'speaking' is often built into the English curriculum exists in most public-owned (e.g. The University of Malaya, University of Science Malaysia,

National University of Malaysia, etc.) and private-owned (e.g. National Power/Electric Board University, University TELEKOM Malaysia, etc.) universities in the country. The speaking component had always been an important feature in the national curriculum where English is taught from year one till the end of the final year in secondary school. However, even though English is commonly and extensively used, Malaysians have recently found themselves grappling with the issue of English language proficiency. This is a national concern as both the government and population realize the need to be IT literate in today's information age, and English is the lingua franca of the Internet. Hence, with technology, the Information age and globalization, the status of the English language in the country heightened. This became evident with the arrival of the MUET (Malaysian University English Test), described below.

B. The MUET exam

The MUET was designed & developed by the Malaysian Examinations Council and consists of all four skills of Reading, Writing, Listening and Speaking (see 'MUET Manual, Malaysian Examinations Council' 2001 for regulations, syllabus & sample questions). It was first introduced in 1999 with the dual purpose of filling the gap with respect to the training and learning of English and that of consolidating and enhancing the language literacy of the Sixth Form and pre-university. The syllabus aims to equip students with the appropriate level of proficiency in English so as to enable them to perform effectively in their academic pursuits at tertiary level. The test instruments are devised to measure the respective skills at a subsidiary level of proficiency, i.e., between the Credit Grade of the O-Levels and the minimum pass Grade of the A-Levels English. At present, it is a compulsory test for all Malaysian

students entering public-owned universities. Since it was introduced, the focus of language classrooms in the country has shifted dramatically from preparing students for language acquisition and language skills to preparing many students for the MUET exams. Schools, universities, and other private institutions alike provided students with reading & practice materials to familiarize them on the test format, topics, content, etc.; the market was quickly filled with MUET-based publications. Leading Matriculation colleges, private English language institutions such as ELS (English Language Services) Malaysia, and other universities have either offered external MUET courses, or incorporated a MUET test preparation course in their programmes. For the listening and speaking components, CDs were produced to provide candidates with relevant text, questions, and practice exercises.

To date, the council has published the MUET test results on an annual basis, and it has included academic personnel from universities and schools around the country in its rating and test development meetings. However, little is known about revisions and/or research work that they may have conducted on the test. According to MUET Coordinator (personal interview with researcher in Jan 2005), there are on-going rater training and item development programmes. Validation has not been conducted for the speaking test, but a correlation study has been conducted for the Writing test, correlating scores from MUET writing with scores from IELTS writing; reports on how other components correlate with each other are not published.

Given its importance and enormous scale throughout the country, the impact of the MUET is prevailing; it has affected teaching and learning in English classrooms, encouraged MUET course materials production, and encouraged parents to enrol

their children into these courses. It seems that the MUET has had an impact on the education system and on university students specifically; students at Mara University of Technology especially are aware of what is expected and required of them in order to pass the exam and achieve the target band for entrance to university.

1.2 Context of the Study

A. Testing of Speaking at the university

As stated earlier, the proposed research on testing spoken language via the computer is directly linked to various centres and programmes at the university. Like the full-time students, students who embark on the distance-learning mode or part-time basis for their courses, have to enrol for at least one English course during the academic year. The EDC for example, offers a language programme to students from various fields of study, such as those in Business Studies, Accountancy, and Law. Students at the centre attend four seminars in a semester, and have the advantage of going on-line with their facilitators, administrators, counsellors, other fellow students, and so on. They have access to e-mail, forum (group, subject & individual) and chat facilities, and other web services. Through the centre, students have access to the English programmes where 'speaking' is a component. The English courses are conducted in the same manner as the other subjects, through four face-to-face meetings/seminars, and through web facilities for materials and interaction with facilitators and other candidates. However, candidates have to attend examinations at the respective campuses, as their full-time counterparts do, and the English exams are no different. The speaking component is still 'taught' and tested in the classroom, using the face-to-face/direct method.

The Language Centre provides the English courses for all faculties and other centres at the university. They range from Proficiency English for diploma candidates, English for Specific Purposes (ESP), English for Occupational Purposes (EOP) and English for Academic Purposes (EAP) for candidates enrolled in a degree or post-graduate level programme. As the centre was involved in the development of the MUET from the start, the Proficiency English curriculum at the university was therefore adapted to suit the MUET. This was done so students not only undertake an English course, they also benefit from a MUET-based syllabus, which directly prepares them for the test. The course covers the four skills and for each skill the test would have similar contextual features as those of the MUET. Consequently, students are prepared not just for the speaking test at the university, but also for a national-level speaking test. In this test, they are expected to present ideas, elaborate on the ideas through explanations, examples, etc. in the target language via two tasks: an individual presentation and a group discussion.

All the English courses have 'speaking' built into them, either as a component or as a main course. For example, speaking is a component in all three proficiency English courses at the diploma level, but at the higher levels, speaking could even be a course by itself, for e.g. "Speech Communication" for a degree in Business Studies. Hence, all students at the university enroll for an English course in which speaking is an important component, and various forms (presentations, discussions, forums, etc) of the test are administered to students throughout the year; the importance of being able to 'speak' in the target language, could not be more evident. For the diploma students, the stake is higher as their English lessons prepare them for a bigger, national level examination.

B. Rationale of the study

The researcher is particularly interested in this topic because of her tenure as Language Coordinator at the Centre for Distance Education (now Education Development Centre/EDC) at Mara University of Technology Malaysia, where distance learners who come from various parts of the country currently gather at one centre to take the direct speaking test, administered by the Language centre. Students in the distance-learning programme can be found in various branch campuses in the other states around the country. The total number of students in these campuses adds to more than 50,000 and every year, about 10,000 of them take an English exam, which will involve a speaking test.

The task of administering a direct speaking test to large numbers of students across many test centres involves many weeks of planning and preparation. Many hours are taken up for test development, preparation, administration, rating process, data analysis and so on. The speaking test is prepared and developed by the Language Centre at the main campus and it is then distributed to all other campuses; it is usually administered throughout a week and parallel forms are used during that week. The question of practicality and feasibility of conducting a test of this nature to thousands of students raises the issue of validity of the assessment on a very large scale; how marker reliability and test validity can be maintained is a critical point.

Given the scope and pressing concerns that the test takes on, it is pertinent that research is conducted so alternatives can be found to address issues that the speaking test poses for the university. An empirical study to validate the existing

speaking test is required so more information can be gathered on the test, and its strength and weaknesses can be determined. Moreover, given the problems of the speaking test in question, the ability to demonstrate the validity of the construct we intend to measure in the test is of utmost importance so as to ensure it does not suffer from 'construct under-representation' or 'construct irrelevance' (see Cook & Campbell 1979; Messick 1992; Heubert &Hauser 1999; Murphy 2004). In this study we will attempt such a validation process on the direct speaking test before we attempt to deliver it using the computer. Consequently, this study will explore the potential use of computer technology as an alternative mode for testing of speaking, in the hope that it will help overcome some or all of the problems the direct method presents.

The literature on testing of speaking (ref Ch 2: section 2.2 on 'Testing of speaking') shows various procedures of testing which have focused on the direct/face-to-face method, the semi-direct (usually using audio or video capability), and computer-based testing of which research on 'speaking' is also limited in the field. The literature also points to the fact that although studies and reviews have been done on these tests in terms of their validity and reliability, the testing literature is still lacking a systematic, comprehensive framework which can be used for validation of these tests. Validating a test has been standard practice for test developers since the early 1950s, but none so far has shown evidence of test validation in clear and systematic terms. (ref Ch 2: section 3.2 on 'Issues of Validity' for details).

To meet this shortcoming, this study will refine and operationalize a new socio-cognitive framework for the test of speaking (related to Weir 2005) to validate the existing speaking test at the university, and to validate the new test to be

administered by the use of the computer. The framework takes into consideration the important elements of context validity, theory-based validity, scoring validity, consequential validity, and criterion-related validity. The study, however, will focus on the first three validity components (context, theory-based, scoring); which together constitute the construct (speaking) we are testing. Further discussion on this point will be found in chapters 2 and 3. It is also beyond the scope of this study to include consequential and especially criterion-related validity in its investigations; in fact, some information had been gathered for these components but was insufficient to be described as contributory factors to these aspects of the validity of the speaking test.

1.3 Organization of the thesis

This chapter has presented the scope, objective and rationale for the study on the validity of testing speaking using computer technology. The major concerns and areas of the study will be on the validation process of the speaking test (both direct and semi direct), the design and development of a speaking test using computer technology, and a comparative analysis on the validity of both methods of testing speaking in order to justify replacing the direct test with a semi direct test of speaking ability.

Chapter 2 will review the literature on the spoken discourse, the testing of spoken language, and on issues of validity. Specifically, it will look at various linguistic and psychological perspective of the spoken discourse; take on a historical perspective of the testing of speaking, and a chronology of the different methods of testing speaking (direct to semi-direct methods, to the use of computer technology for

computer-based tests). References are made to frameworks that have developed progressively on the testing of speaking and how performance is affected by various factors systematically (McNamara 1996, Skehan 1998b, O'Sullivan 2000), on importance of contextual features in L2 use (Chalhoub-Deville 2000, 2003, Young 2002), and on other facets of the testing of speaking (McNamara 1997, Hughes 2002, Fulcher 2003, Luoma 2004, Dimitrova-Galaczi 2004, Brown 2004). At this point, the literature on the testing of speaking using computer technology (CBT, CAT, and WBT) though scarce is presented. It is an important section of the thesis as the case is made for the advantages of testing speaking using the computer; for the Malaysian context especially, standardization of administration and reliability can be addressed. As the study shows, important issues such as co-construction of discourse (see Fulcher 2003, Luoma 2004, Dimitrova-Galaczi 2004) may have recently surfaced but has been present in most if not all oral tests that involve interaction. A computer-based oral test offers technical and procedural innovations in terms of efficiency in administration, scoring and even examinees' affective responses (Kenyon & Malabonga 2001). Norris (2001) however, argues that a more fundamental question to ask is not "how" to use it, but under what circumstances and for what purposes should we use it. He stresses that a comprehensive programme of validation is long overdue on tests or guidelines (e.g. ACTFL Guidelines), which are used as the basis for developing computer-based tests (see documents on the COPI: Kenyon & Malabonga 1999). Most importantly, we have to be sure whether we are able to make warranted interpretations about examinees' knowledge and abilities based on critical features of speaking performance that the computer captures.

Test validation is a major part of the study, and the chapter discusses validity in terms of the different types of validity, and how the construct 'speaking' and the speaking test has been validated to date. While formal test validation has been in existence since the 1950s, it is clear that an empirical, systematic and comprehensive study, using clear guidelines or a framework, is missing in the literature. Hence, the framework for validating the speaking test (Weir 2004/05) is introduced as it is used throughout the study. The framework for validating skills in language tests is presented with more detailed descriptions, justification for its use, and the focus on the framework for validating a speaking test is discussed. The chapter then leads to the research questions for the study, which, accordingly referred to the three major concerns of the study: test validation, development of a new test and its validation, and comparative study of the two test methods.

Chapter 3 is on methodology of the study. This chapter describes how the validation framework acted as an 'anchor' for the study as it was used from the beginning to the end; it was used in the development of survey instruments (questionnaire and semi-structured interview), for the validation studies on the direct test and the computer test, and the final comparative analysis of the two testing methods. It is divided into four sections. *Section I* provides a detailed 'theoretical' model of the study as it was conceived when the study first began, and this was organised according to the various stages that the study will undertake to address the research questions that were formulated at the end of Chapter 2. *Section II* is a detailed account of the development of the research instruments used in the study which include questionnaires, interview guidelines, observation and participation schedules; pilot studies which were conducted at various stages of the study to try/test the instruments are included in this section. Essentially, instruments were developed for two main studies (direct test and computer test); in addition to the instruments listed

above, work related to the computer test involved developing the test itself and test script, test specifications, and the adoption of a new rating scale. *Section III* contains a description of participants and the locations involved in the study, and a matrix is presented which shows the instruments used at each stage of the study. Finally, *Section IV* presents the model of the research plan which illustrates the actual conduct of the study which involves five phases. This section illustrates and describes the data collection process and procedures that were actually carried out; and hence it was termed the 'operational model' of the study.

Chapter 4 provides details on how data were gathered from the various pilot studies and main validation study for the direct speaking test. It includes detailed information on the findings of the study in terms of the various sources from which they were gathered, including conclusions for each aspect of validity and overall conclusions of the findings for the direct test validation. All quantitative data from the questionnaire surveys were organized and analysed in SPSS software while qualitative data gathered from interviews were organized and analysed in Hyper Research software.

Chapter 5 provides details on how data were gathered from the various trials, a pilot study and main validation study for the computer-based speaking test. It includes detailed information on the findings of the study in terms of the various sources from which they were gathered, including conclusions for each aspect of validity and overall conclusions of the findings for the computer-based test validation.

Chapter 6 is the chapter on a comparative analysis of the outcomes of the validation studies of the two tests. It illustrates the differences (and similarities) between the tests in terms of context, theory-based and scoring validity.

Chapter 7 concludes the thesis with a summary of the major findings of the study in relation to the research questions. The contributions of the study are described in terms of the Malaysian context specifically, and research on ESL/EFL language learning and testing in general. The limitations of the study are highlighted and several recommendations for further research and investigation are projected.

Chapter 2 LITERATURE REVIEW

2.1 INTRODUCTION

As stated in chapter 1, this research is on the testing of spoken language using computer technology. The objectives, rationale and scope of the research were described in that chapter; however, it is noted that developing a brand new test using advanced computer technology is not the main focus of the study. What is paramount is to examine the issues surrounding the testing of spoken language as it is carried out today, using a structured validation framework to obtain data from an existing test, and eventually introducing the speaking test on a new platform (semi-direct and computer-delivered) in an attempt to address the drawbacks and problems that the direct testing method encounters. Therefore, this thesis ultimately aims to provide a clear, systematic basis to its practical and theoretical outcomes.

To date, the literature on testing speaking shows that changes have taken place over the years since the 'live' testing methods such as the oral proficiency interview (OPI) in which the candidate interacts in the target language either with an interviewer or with another candidate (or both), for example when the Foreign Services Institute (FSI) examinations (1956) was the leading indicator of oral ability in the USA. It is now commonly accepted (Lazaraton 1992; 1996a; 1996b; Shohamy 1983) that different test formats or interaction types such as role plays, discussions, and presentations are necessary for construct validity of a test. Accordingly, in many current oral tests the candidate engages in a number of different interaction types, which offer a broader view of the overall speaking ability (and hence a more valid assessment of a learner's ability to use the language in terms of both content coverage and theory-based

validity). In addition, other less direct or semi-direct methods of testing speaking have been produced in an attempt to address concerns which the direct method of testing faces (see Lowe & Clifford 1980; Clarke 1979; Stansfield & Kenyon 1992 for a fuller description and discussion of the simulated OPI/SOPI in which input is tape or video delivered; see Brown 1997; Chapelle 2001; Fulcher 2000; Goodwin-Jones 2000/2001; Norris 2001; for issues surrounding computer-based testing; Jones 2001, on web-based test, Kenyon & Malabonga 1999, 2000, 2001, on CAT for speaking, on-going research projects at UCLA (Bachman et al, presented at LTRC 2002), and the DIALANG project (Alderson et al 1998).

It appears from the above list that there is a possibility that with the advance of technology in education, especially for learning and testing, the present study is able to take the best features of technology to deliver the speaking test while addressing concerns of the direct test method. In other words, as discussed in the sections below on computerized testing (section 2.123) and validity (section 2.2), we can take out the best from existing direct tests in terms of content, theory-based validity and add the positive elements of the computer-based test in terms of reliability and practicality. In order to achieve this, we need to investigate several issues that literature in spoken discourse and on the testing of speaking have raised over many years, i.e. the process of speaking & factors that influence this, especially where interaction is involved such as in natural conversation or in more formal discussions, and how all these features are reflected in our test so that inferences we make of test scores are accurate and point to a candidate's speaking ability and nothing else, so far as possible.

This chapter will consist of the following major topics:

◊ Literature on Speaking

◊ Literature on Computerized testing

◊ Literature on Validity

The section on speaking includes discussions on the theories of spoken discourse, i.e. what it involves, what factors influence a person's speech, and why in spite of its significance to human life, the literature still lacks explicit attention on speaking as ability in its own right. Hughes (2002) claims this is due to several reasons such as the idea that it is not a discrete skill, that teaching it is not easily separated from other objectives in the language classroom, that there is a distinction between teaching speaking and using it to teach language; however, recent work on corpora of natural speech and language processing (see Carter & McCarthy 1995/1997; Carter, Hughes & McCarthy 2000) opens other avenues for further research in the field. It is important to survey these areas for a better perspective of what 'speaking' means before we can develop a test which is well-founded in its theoretical underpinning of the construct we are attempting to measure.

This leads the chapter into the discussion on the testing of spoken language. This will include areas for consideration for testing which involve aspects of the test task, the test taker, interlocutor, rater and issues related to the rating process, and methods of testing speaking. In the last section of the chapter we look at the potentials of a computer-delivered speaking test.

The next major section on validity looks into topics such as what is validity, why validate, validation in language testing, types of validity evidence, and validating

the construct in a speaking test. Areas related to the process of validation are central for any research that involves testing and the use of tests. In the present study, an existing direct speaking test is used and analysed to establish its construct validity (see Weir 2004 & discussion below on validity for details) by looking at specific elements of the test in terms of the test task(s), the test taker, and how the test is scored. From the review of the literature, it is hoped that several questions will emerge that will direct this study into its investigations.

2.2 Spoken discourse

The importance of spoken discourse or speaking for various purposes in various domains has long been recognised. Hughes (2002) cites how Roman civilization and scholars of 100BC perfected the theories of oratory from the Greeks before them; the emphasis then was on the ability to speak for the purposes of persuasion and rhetoric. This emphasis on teaching speech as the basis of rhetorical devices changed towards an emphasis on speech as it is used in the education system. Even though the move from the 'grammar translation' approach of the late 19th century Europe was replaced by the preference for more 'natural' and 'direct methods, and later the communicative approaches of the 20th century, Hughes (2002: 23-25) contends that the spoken language as a skill in its own right was under-investigated and was regarded as less important than writing. Bygate (2003) echoes this concern, stating that speaking is usually used in the classroom for purposes other than developing the skill of speaking or oral discourse, but rather to model language or highlight accuracy.

In the field of language testing, although the focus on testing second language speaking did not really surface until World War 2 (Fulcher 1997), the Cambridge

ESOL Certificate of Proficiency in English (CPE) offered in 1913 has by far the longest track record of any EFL exams still in existence which has an oral test as a component (see Weir & Milanovic 2003 for a review on the history of testing in the UK).

Until then, in the USA, the 'oral test' involved pronunciation and remained mainly paper and pencil tests. This is because of the heavy emphasis on the reliability of test scores, subjectivity of such tests and logistical difficulties (see Spolsky 1995 for a review on the history of testing in the USA). In 1950s the first test of speaking skills with a specifically designed rating scale were developed for the FSI (Foreign Service Institute) personnel as evidence of their language ability while carrying out duties abroad. The oral ability of language learners was assessed during a 'live' interview, in which the learner interacted with a lone examiner, resulting in an 'interview' type discourse being produced (O'Sullivan 2002/2004). The FSI approach to testing speaking became popular and was adapted for use by other government agencies and universities nationwide. Later the ILR/Interagency Language Roundtable (1960s) and ACTFL/American Council on the Teaching of Foreign Languages (1982) produced revised versions of the standards used for rating the test. In the UK, although the testing of speaking also had a military focus, the major concern was in defining and coordinating standards for examiners throughout the country (Spolsky 1990, 1995; Weir 2003, 2005; Fulcher 2003).

Hence, in the early history of testing speaking, there was a focus in the USA on the development of rating scales for the speaking test because this would affect the scores awarded to candidates, which in turn affects test reliability (in Fulcher 2003).

In the UK, however, the focus had been on the validity of the construct that is measured, i.e. speaking (see section 2.12 below for details on the history of testing speaking). There was a concern for validity but heavy reliance on judgements rather than rating scales. The CPE papers of 1913, and its future revisions were examples of this. The oral tests consisted of reading aloud & conversation (30 minutes) and associated dictation (30 minutes); this format remained very much the same right through to the 1960s, with no mention of how the test was assessed. Changes in format & assessment only became evident in the 1975 revisions when the oral test was changed to a series of interviews based on different inputs such as a picture or a topic; assessment was based on a range of criteria which then translated to an overall level of communication.

Therefore, while emphasis on the importance of spoken discourse had been going on for centuries, it was only in the second half of the 20th century that further research on its actual nature & process, especially in conversation and interaction, and factors that affect or influence speech patterns were conducted, and theories formulated.

2.2.1 Theories on spoken language

2.2.1.1 In the field of Psycholinguistics

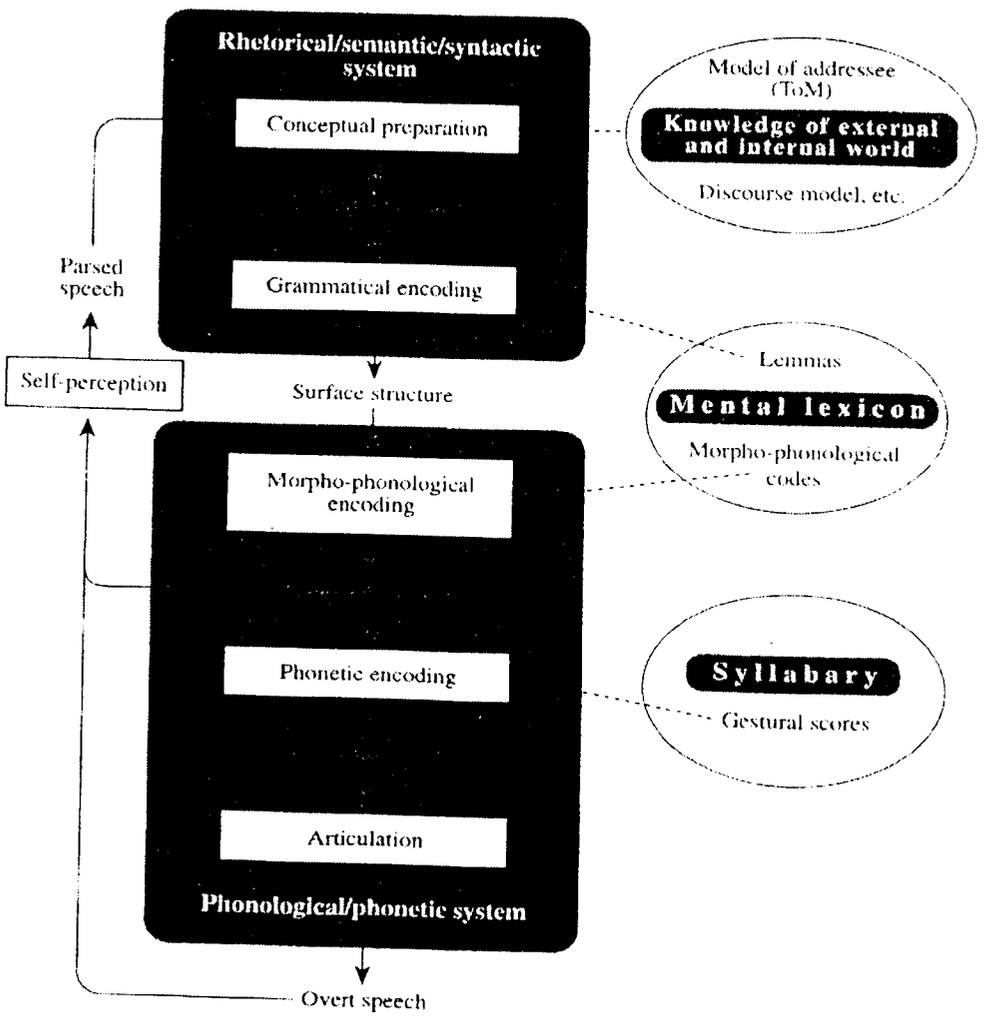
In the 1970s, studies in spoken language und erstanding and processing were mostly found in psycholinguistics. There were two common views: that processing is sequential and each component is autonom ous in its operations, and that processing is a more flexibly structured system (cf. Fodor et al. 1974; Tyler & Marslen-Wilson 1977; Marslen-Wilson et al. 1978; Marslen-Wilson & Tyler 1980; Carroll et al., 1978;

Forster 1974, 1976, 1979). However, the primary concern for psycholinguists was in fact how spoken language data relate to underlying linguistic systems.

Levelt (1989) was the first to propose a model which explains the processing system that underlies speech production. The model illustrates the organization of the speech process, from the constraints on conversational appropriateness to articulation and self-monitoring; a more comprehensive system than theorised and perceived before. Seeing the speaker as an information processor, Levelt proposes a blueprint (see diagram below) in which message generation, grammatical encoding, phonological encoding, and articulation are seen as relatively autonomous processors. Two points are however, made clear. First, though Levelt's model stops largely at the point of utterance, he devotes an entire chapter to the speaker as an interlocutor in natural conversation (Chapter 2). Here he describes at length the three essential aspects of conversation in which the speaker is a participant and interlocutor: it is highly contextualized, has spatio-temporal setting, and is purposeful. Secondly, though the model may seem complex at first, the basic mechanisms of speech processing are conceptualized in a fairly basic manner: we produce speech by first conceptualizing the message, then formulating its language representation (encoding) and finally articulating it. With reference to speech perception, speech is first perceived by an acoustic-phonetic processor, then linguistic encoding in the speech comprehension system (the parser), and it is finally interpreted by the conceptualizer. A detailed account of how speech evolves in humans from infancy to adulthood, and the blueprint to illustrate this mechanism is found in Levelt's chapter on "Producing spoken language: a blueprint of the speaker" (see Brown & Hagoort 1999).

The following diagram illustrates the speech process according to Levelt (1989):

Figure 2.1 A blueprint for the speaker (Levelt 1989)



Levelt's work in terms of the blueprint/framework influenced other researchers and has been used in works by or referenced in more recent works on research in speaking (see Dornyei & Kormos 1998; Ortega 1999; Bortfeld et al. 2001; Hughes 2002; Ellis 2003; Weir 2005).

In the present study, the blueprint forms the foundation for theory-based validity/internal processing component of the framework for validating a speaking

test (Weir 2005). This aspect of the validity framework is essential, not just for the purpose of validation but also for a better understanding of the processes or operations that test takers utilize when attempting the test task; only through such data can we make decisions about these operations in relation to the elements we include in the test task to ensure context validity. Furthermore, since literature on L2 learners' problem-solving strategies in communication is limited to research conducted in the area of linguistics (such as Cohen & Olshtain 1993; Cohen 1996; Cohen 1998; Gass 1997; Dornyei 1995), it is imperative that we investigate how these strategies affect and apply to performance in language tests.

2.2.1.2 In the field of Linguistics and Applied Linguistics

Hughes (2002) maintained that although spoken language has been highly valued in linguistics and applied linguistics, there is still little attention paid to how the spoken form is viewed in its own right. This is partly due to difficulties in obtaining & classifying speech samples, unlike samples available for the written form. Hence, analyses had been on speech at the discourse level with focus on 'rule-based' theory looking into rules of interaction and structures in discourse. Moreover, in language teaching, teaching speaking is often linked to other underlying language acquisition objectives such as cultural difference awareness and conversational appropriacy in context (see Hargreaves & Fletcher 1979; Brown & Yule 1983; McCarthy & Carter 1994; Ur 1996; Carter & McCarthy 1997).

One of the basic aspects of spontaneous speech is that it happens under real-time processing constraints (Hughes 2002: 135). Among others, studies on second language learners' management of the target language in monologue and interaction

have a focus on the processes that learners acquire and utilize to overcome their speech difficulties. Gass (1997) presented a framework for second language acquisition in which input, supported by awareness and attention, and mediated by other factors like time pressure, frequency of input, and prior knowledge, determines the quality of language output. More importantly, interaction in conversation encourages negotiation of meaning and form and this promotes comprehension and management of input among speakers. Dornyei & Kormos (1998) take a psycholinguistic approach to show that learners' problem-solving mechanisms at various levels of speech production can be linked to consecutive stages of speech processing as described in Levelt's framework. This is parallel to Skehan's (1998) reference to VanPatten's (1990) research which proposed that when most language learners are confronted by difficulties, processing is a question of priority between form and meaning; usually meaning will take preference as fewer resources will be available to attend to form.

It is clear from the above studies that language learners may employ some form of internal processing when they are acquiring a language; whether they attend to form or meaning first, a process is taking place within the learner. This then applies to the test context where candidates have to attend to the task; in the process they draw upon resources element (content & linguistic), a processing component and a monitoring component through which the former components interact to produce a response (Weir 2005).

2.2.1.3 Other works on spoken discourse

Earlier studies which attempted to describe 'speaking' according to its process and

functions emerged from the need for a clear definition of the construct, i.e. what it means to be able to speak in various context and how speaking is measured.

Communicative competence

Ellis (1980) drawing on Halliday's (1975) four categories of language functions, emphasized the importance of making learners aware of these functions, at a time when teaching oral skills was significant for students to attain "life skills", for success outside of the school setting. His concern was related to the teaching of oral skills when aims for oral work were at the time wide and vague and proposed a checklist of oral skills that could be used for teaching the skill. Other works attempted to describe the spoken discourse based on what it means to be competent or proficient in the language (Filmore 1979; Adams 1980; Lantolf & Frawley 1988; among others). In addition, many more works were found on what communicative competence means, especially sparked by Chomsky's (1965; 1976; 1980) concept of language competence; these include Hymes (1972; 1985), Canale & Swain (1980), Canale (1988), Spolsky (1989), and Widdowson (1989), among others. For example, Widdowson (1989), argued that communicative competence is not just a matter of having knowledge of abstract systems (Chomsky' & Hymes' beliefs) and the ability to employ the rules to assemble expressions (Hymes' belief). Rather, he was of the view that competence is knowing a set of pre-assembled units and the ability to apply rules where necessary to fit particular contexts (Widdowson 1989: 135). Hence, association between lexical clusters & context without need for grammatical adjustments is sufficient, when this becomes insufficient to establish meaning, grammatical rules are called in to assist lexis with syntactic fittings required.

Beyond clarifying the notion of competence especially for its importance in language learning and teaching, other works surfaced on how language proficiency is measured.

For example, Lantolf and Frawley (1985) asserted how the construct 'speaking' had been misrepresented by describing proficiency directly and indirectly via psychometric principles. This includes measuring proficiency according to scales & institutionalizing ACTFL/ETS/Interagency (AEI) Guidelines (1986) as criteria to define communicative competence in curriculum guides, and as entrance & exit requirements for university & school programmes. Canale (1988) stated how problems in measuring communicative competence, especially when characterizing its development, have surfaced; for e.g. ACTFL has come under much criticism mainly due to the problem of identifying plateau points (stages in development at which further development may be difficult). Another concern is how relevant information on measure of production skills is gathered, such as the ILR/ACTFL/ETS for assessing oral interaction (Byrnes & Canale 1987), and greater emphasis is needed on criterion-referenced as opposed to norm-referenced testing. On how to use evaluation information and ethical considerations, he cites Spolsky (1978) on three major historical trends. For Canale however, a naturalistic- ethical approach is imperative; we need to make measures of communicative competence more rewarding, unintrusive and naturalistic, and we need to be more responsible and ethical in our use of information gathered from or about tests.

All these studies point to the fact that we have moved from various perspectives of what comprises communicative competence and how language is viewed. We started from an emphasis on knowledge of how language is structured (syntax & grammar, morphology, lexical/vocabulary, phonology) and how it is used in communication (discourse management/competence), to concerns on how communicative competence is measured. More recent studies and research classify speaking in terms of how speakers attend to and manage communication in interaction of natural conversation.

Bygate (1987) was one of the earlier proponents of a communicative approach to language learning who described how speakers organize their speech according to routines; using knowledge (executive resource) and motor-perceptive skill (processing) to achieve communication and to satisfy particular demands. Processing conditions are affected by aspects of reciprocity, ability to produce speech at normal speed under pressure of time (see Brown & Yule 1983), language facilitation devices and compensation. He identified interaction skills as routines (from Widdowson 1983) which consist of information (expository & evaluative) and interaction, with attention on negotiation of meaning and management of discourse. For Bygate, it is important for learners to be familiar with these features in order to be able to learn to speak and be "normal" in their speech. He also highlighted the differences between speech and writing especially in terms of processing conditions, which is the key consideration in teaching spoken language for the right reasons and using appropriate techniques so these reasons can be achieved (Bygate 2003).

O'Sullivan, Saville & Weir (2002) expanded on this to develop a checklist of functions for validating speaking test tasks. The checklist contains functions representing the construct of spoken ability and is used to match intended and actual test-taker language and content. It consists of informational functions, interactional functions and those of interaction management functions generated by test takers. Being able to match content that is intended for test takers to produce or elicit in a test and what they actually produce, is valuable data for investigating construct validity; defining the construct is underpinned by establishing the true content of tasks (O'Sullivan, Saville & Weir 2002: 46).

In the present research, one of the facets investigated during validation of the direct speaking test is the rating process. The checklist on speech functions as proposed by

(O'Sullivan, Saville & Weir 2002) was used to ascertain if elements of interaction such as conversational repair and negotiation of meaning are evident when students participate in the group discussion (task B of the test; see Chapter 4 for details of the direct speaking test). This needs to be determined first before the rating process, so raters know exactly what to expect and look for in the group discussion task when students interact with each other.

Speaking in natural settings

Several other works began to explore spoken discourse in a more natural setting such as in conversation. Cheepen & Monaghan (1990) details how spoken dialogue in conversation is analysed, with particular attention on discourse management of the speaker called 'speaker behaviour', i.e. how speakers progress through successive utterances to construct meaning and gradually become aware of one another's views, and how interaction breaks down but is handled by repair strategies. They looked at job interviews as an example of the combination of interactional activity within a transactional framework. Hoey (1992) showed how naturally occurring dialogue differs from classroom constructed dialogue in eight respects, and how important it is for language teachers to know how casual/normal conversation is organised in order to better assess a learner's success. Cohen & Olshtain (1993) focused on how a group of non-native speakers of English assess, plan and execute speech utterances through speech acts such as apologies, requests and complaints; data was gathered through retrospective verbal report protocols. The major points that surfaced were the language(s) used when planning and executing speech act utterances, the degree of specificity during assessment & planning, and how delivery reflected very few or none at all of the strategies that were listed in the literature such as use of formulaic speech, monitoring and lexical avoidance or simplification. Riggenbach (1998) looked at how natural, unstructured conversations can provide valuable information on micro-

skills/functions of speech that contributes to learners' discourse, strategic competence and interactional skills. This includes, for example, the ability to claim turns, maintain turns, to backchannel (discourse competence) to self-repair, to ensure comprehension on the part of listener, to initiate repair, and to employ compensatory strategies such as word coinage, circumlocution and shifting topics (strategic competence). Kormos (1999) investigated the nature of interaction in two oral proficiency tasks of non-scripted interviews and guided role-play activities. This is to examine how the type of task candidates take part in affects conversational structure and the candidates' opportunities to display their conversational competence. The findings suggest that guided role plays can be better measures of a candidate's ability to manage conversation, i.e. performing openings & closings, initiating & rejecting topics and interrupting. Pridham (2001) also described naturally occurring spoken language in conversation as spontaneous, temporary unless recorded, and potentially taking place through body language, prosodic features, & even laughter or silence. The 'rules' of conversation include turn taking/adjacency pairs, the ability to negotiate, with purpose, and predictability (some predictable: daily, repeated, some less predictable but frequent: friends telling stories, some structured: in classroom between student & teacher).

In summary, studies on spoken discourse in terms of its description, classification, structural and discourse features, processing, and interactive features have clearly progressed. It has moved from the early theories of the structuralists, psychometricians, and the communicative approach, to more current research that attempts to investigate spoken language as it is being used in more natural settings, such as in a conversation and to shed light on what this type of speech communication really involves. (see Hughes 1996, 2002, Tannen 1989/96, Gaskell & Marslen-Wilson 1999, Pica et al 1996, McNamara 1997, Johnson & Tyler 1998, Johnson 2001, Carter & McCarthy 1997, McCarthy 1998,

Mackey, Gass & McDonough 2000; on spoken vs. written discourse see Olson 1977, Hughes 1996, Biber et al 2002; on measuring units of spoken discourse see Crookes 1990, Foster & Skehan 1996, Ellis 1996, Lennon 1990, Tonkyn 1996,Wigglesworth 1997,Willis 2000, Foster, Tonkyn & Wigglesworth 2000).

Findings from these studies have serious implications on how we perceive our tests of spoken language to be, i.e. the nature of speaking tests or assessments that we develop. We need to be clear about where the linguistic resources available to the candidate are consciously reflected in the test task and are again reflected in the rating criteria (O'Sullivan 2004). We need to incorporate elements of linguistics (structural, lexical, functional), elements of the task setting such as a clear purpose and other demands of the task such as response format and length of test into the contextual features of the test. We then consider the processes (cognitive & meta cognitive) that candidates have to employ in attempting the task given their background & knowledge (language & content) of the task, and finally elements of scoring or how the test will be assessed such as rating criteria, the rater, the rating process and conditions.

The next section on the testing of spoken language will highlight these elements in the speaking test and how they are validated to ensure we have produced a reasonably fair and credible test of the ability we want to measure.

2.3 Testing the spoken language

Hughes (2002, ch.4) reminds us that the very nature of speech makes its potential for subjectivity, variation in test facets, and maintaining consistency across tests higher than those of other language tests. In addition, research has led us to question how far the test

is a test of speaking, as opposed to testing more general language proficiency; and if it is possible to incorporate characteristics of natural spoken discourse into assessment models that exist today.

This section looks at the various potential procedures for testing oral proficiency in terms of their characteristics and contributions, historical development, and the problems arising from their use.

2.3.1 The direct speaking test

For a long time, direct speaking tests, like the oral interview, have been the industry standard. In fact, they have a long history that dates back to the 17th century in the USA. In the United States, Harvard University had a 1650 statute that required an oral examination (Buck 1964, in Spolsky 1995, p.21). In 18th century Europe, Latin was the language most studied by young men who wanted to enter prestigious universities such as Oxford and Cambridge. New methods in examining spoken Latin involved speaking activities which raised the issue of the backwash effects of the test in the classroom (James 1983). These tests, however, were delivered orally (and responded to orally) through the medium of language; they were not necessarily tests of language per se.

In Britain, the University of Cambridge introduced its first test of English for foreign students in 1913 for the Certificate of Proficiency in English (CPE); it included an oral test of reading aloud and conversation, with associated dictation (Weir & Milanovic 2003). Thus, the CPE had an oral component at its early stages of conception, and this format remained well into the 60s when the CPE was revised further, and an oral interview task was included. After three further major revisions (1975, 1984 and

2002), the oral paper now has three parts which involve an interview, a discussion, and a long turn followed by a discussion.

Throughout the 20th century, other oral tests were instituted elsewhere in the world, such as: the French oral test in Canada (Ferguson 1930), the Yale Spanish oral interviews (Hall 1932), the oral test of the Army Specialized Training Programme (ASTP) (Angiolillo 1947), oral production test in the College Entrance Examination Board (CEEB) (1954), the Modern Language Association (MLA) oral test of foreign language (Brooks 1960), and the Test of Spoken English (TSE) (ETS 1969) (all in Spolsky 1995, ch.5)

Most empirical research on direct oral testing, however, had focused on the US Foreign Services Institute (FSI) oral interview test developed in the 1950s, later called the Oral Proficiency Interview (OPI), which is widely seen as the first modern speaking test. In the OPI, the examinee converses face-to-face with one or two trained testers on a variety of topics for ten to thirty minutes. The elicited sample is then rated, in global terms, using the Interagency Language Roundtable (ILR)/ American Council on the Teaching of Foreign Languages (ACTFL) scales, which start from zero for no functional ability to a maximum of five for proficiency equivalent to that of a well-educated native speaker.

Claims of the oral interview

Although the direct testing of speaking has had a long history, the oral interview test is still the most widely used oral test. The first reason for this is its claims for reliability associated with high degrees of inter-rater consistency, as found by Adams (1978), Mullen (1978), Clark & Swinton (1980), Hendricks et al.(1980),

Shohamy (1983), Morrison & Lee (1985), Shohamy, Reeves & Bejarano (1986), and Barnwell (1989), among others. The second reason is the claims for validity in its authenticity, directness, or closeness to real life non-test situations, as found by Clark (1979), Wilds (1975), Madsen & Jones (1981), Lowe (1983), Gasparro (1986), and Henning (1987). These studies have found the oral interview test to be reliable and valid in terms of its procedure, rating and content.

Since 1982, even more uses of the OPI have surfaced. In addition to traditional uses at universities as graduation requirements and as part of investigations of the effectiveness of study abroad programmes, some high schools used it as a requirement for graduation. It has also been used for both testing and programme restructuring purposes (see Malone 1999: 53-55).

While the major advantage of the direct speaking test is its characteristics that enables the candidate to speak directly to an interlocutor, and to interact with an interviewer, these points have also been seen as problems with the test. For example, numerous objections have been raised about the OPI in its claim that it constitutes a conversation (Lantolf & Frawley, 1988; Van Lier, 1989; Ross & Berwick, 1992; Young & Milanovic, 1992; Lazaraton, 1992, 1996; Johnson, 2001). Most of these studies show that the interview is not synonymous with a conversation. It does share features with a conversation, but they are still characteristically instances of interviews. Other objections are made in relation to:

- its claim of high inter-rater reliability when in fact, interviewer variability, particularly in the unstandardised OPI, threatens the reliability of the test (Underhill 1987, Hughes 1989, Lazaraton 1996a)

- the validity of the ACTFL guidelines due to the experiential nature of its development, lack of empirical studies conducted on the Guidelines, and the lack of construct validity (Bachman & Savignon 1986; Lantolf & Frawley 1985; Clark 1988; Meredith 1990; Barnwell 1996)
- its practicality in terms of administration and cost. It is often difficult and impractical to organize a direct test which involves many test takers, interlocutors, and examiners/raters, especially if it is taking place in many different locations. High cost is incurred for the selection and training of interviewers and assessors, and for the overall administration of the test (Underhill 1987; Malone 1999; O'Loughlin 2001).

Hence, although language educators and testers have used the test of oral proficiency in the form of the direct speaking test for centuries, questions remain about its validity, reliability and practicality.

The group interaction

However, another format of oral testing i.e. the group interaction task has already been discussed and seen by experts as an important format that can promote language learning and ability. The discussions centred around the nature of conversation and interaction, factors affecting such performance, and strategies for communication in group interaction. In addition to the studies mentioned above on various works related to the spoken discourse, works by Zimmerman & West (1973) drawing on work by Sacks, Schegloff & Jefferson (1974) on turn taking models, and Beattie (1981) for example, discussed how sex and status of interactants are major factors that influence conversation in and among groups; this is evident during interruptions, lapses in flow of conversation, inattentiveness and turn-taking.

Long & Porter (1985) and Long (1989) discussed at length how group work as opposed to the teacher-led, "lockstep" mode can help promote second language performance and acquisition for non-native speakers. In addition to pedagogical arguments for the use of group work to promote language acquisition, a psycholinguistic rationale emerged especially from research on conversation between non-native speakers (NNSs), or interlanguage talk. Further works showed parallel findings that NNS negotiate meaning and conduct problem-solving tasks better when there is interaction, especially with other NNS as opposed to interacting with native speakers (NSs). For example, Gass & Varonis (1985) found that group interaction among NNs promotes meaning negotiation which in turn allows the conversation to proceed with minimum confusion because speakers are allowed to 'manipulate' input. Duff (1986) contended that pedagogic tasks of various types, especially creative problem-solving type tasks, promote interaction among NNSs in that they require learners to make use of world knowledge and past experience, both linguistic and non-linguistic. Rulon & McCreary (1986) provided preliminary findings on how small-group interaction encourages significantly more negotiation of content amongst learners, which in turn is essential to the promotion of interaction necessary for successful second language acquisition.

Similarly, Porter (1986) drew on work by Long (1981) and later Pica et al (1996) on input and interaction features to show how input is important in determining how learners talk to each other in task-centred discussions. She found no clear advantage for a native speaker as an input provider but that second language learners will derive greater benefit from talking to other learners of equal or higher proficiency level, and that learners also need to be exposed to forms and strategies necessary to develop sociolinguistic competence.

These studies showed how group tasks can be effective in promoting language performance because learners can demonstrate how to negotiate meaning in order to solve the problem at hand via a discussion or in conversation with other speakers. They also highlight the importance of interlanguage talk and how L2 learners can benefit most from interacting with other L2 speakers rather than native speakers of the language concerned.

Several works were also established on internal processing of speakers during interaction. In terms of establishing the link between input, interaction and acquisition, SLA researchers have employed introspective methods (from Færch & Kasper 1987) to explore learners' cognitive processes during the task. Among others, Gass & Mackey (2000) and Mackey, Gass & McDonough (2000) used stimulated recall where learners are asked to articulate their thoughts on feedback during interaction and this was analysed for accuracy in their perceptions; this helps the researcher to establish if feedback had indeed helped learners to conduct the discussion appropriately.

Riggenbach (1998) however, looked at audio-taped conversations to establish learners' oral proficiency, and micro skills that contribute to learners' discourse and strategic competence. She established that speech in natural conversation is different from speech in other genres or settings, and it is very unlike claims about native speakers' skills. It is filled with fragments, exchanges in elements of skills to keep the conversation coherent and avoid communication breakdown; its implication for rating is critical as raters tend to look for literate features of speech when assessing speech in interaction rather than the more 'real' characteristics. (see Hughes R.

(forthcoming) for details on literate biases in oral language testing). These studies and others of similar vein (such as Long 1985, Cohen 1996/98, McNamara 1997, Green 1998, Hughes 2002) relate to theory-based validity, i.e. how the test taker who has inherent characteristics and resources, attend to the test task(s) by means of cognitive processing to produce a spoken performance. This need to recognize the internal processes as well as strategies that test takers employ when attempting a test task is relevant in most if not all test validation studies. In the present research, this component is found in the framework as 'theory-based validity' in which data was gathered to ascertain whether and how students go through cognitive processing in order to fulfil the assumptions that test developers, writers, experts etc. make about their ability and knowledge of spoken language.

Group task in language testing

In terms of language testing, studies in paired or group oral testing focused on how many more variables need to be taken into consideration when such tests are conducted (Berry 1994, 96; Fulcher 1996b; McNamara 1997; Foot 1999; Bonk & Ockey 2003; Brown 2003/2004; Chalhoub-Deville 2003; Riggenbach 1998). How candidates perceive the interlocutor in terms of characteristics such as age, gender, and acquaintanceship may in fact affect their performance. Work conducted at Reading University related to affective reactions of the test taker to his/her interlocutor such as O'Sullivan (1995) & O'Sullivan & Porter (1995) on age; Porter (1991a, 1991b), Porter & Shen (1991) on gender and status ; O'Sullivan (2000a, 2000b, 2002) on acquaintanceship. All these studies were outcomes of what Porter (1991a) referred to as 'affective factors' which may be responsible for a significant portion of the variability in candidate's test performance, while McNamara (1996,1997) and

O'Sullivan (2002) see test performance as being affected by factors related to the test-taker, the interlocutor and the task.

This very point is in fact crucial to the present research in which the speaking test in question (the UiTM speaking test) involves a group discussion between four candidates as one of its tasks. The fundamental concern here is whether the group discussion does indeed enable students to demonstrate functions involved in managing interaction such as negotiation of meaning, conversational repair, agenda management and so on (see O'Sullivan, Saville & Weir 2002 for a checklist of speech functions). If it were the case that in the discussion task candidates produced minimal or no interactional speech functions, then the contextual features of the task need to be re-examined, especially since it is a theory-based validity concern.

More specifically, language testers have considered discussion (in pairs or group) as a popular method of assessing spoken language. Berry (1993/94) in her studies on how individual differences in personality affect performance showed evidence of this, plus how a method effect of the group oral may have influenced how introverts and extroverts perform. She went on to show this further (Berry 1996) in her study on how learners score differently on a paired test of oral ability depending on the test tasks and homogeneity or heterogeneity of the pairs of learners. What was evident was that the degree of interactivity required of the task (e.g. individual or paired task, homogeneous or heterogeneous pairs) determined a learner's speech production in terms of quantity and quality, and this then influenced how raters awarded the scores. Fulcher (1996b) stated the importance of data from affective responses of students to different tasks types in test design, but reminded us of how researchers are responsible for looking into other possible confounding variables

such as rating criteria/scale, before we can make generalizations of test scores from one task to another. More profoundly, McNamara (1997) detailed an alternative to how we view performance in second language performance assessment; we need a broader view, which focuses on the 'social dimension' of interaction in a test. We need to have a better understanding of what happens at the local level during the test interaction itself and not constantly trying to mirror its outcome to real world situation and performance; candidates' interactions are socially constructed and they do not perform in isolation of each other. The rating conditions such as raters' reactions to other interlocutors and the test task, and the rating criteria used for rating, all determine the likelihood of a candidate getting a particular rating. Hence, in general, not only elicitation but also the interpretation of performance is a social act (McNamara 1997: 453, citing O'Loughlin '97 and Berry '94 on issues related to the context of group oral tests).

Studies that analysed components of the group oral task empirically are found in Fulcher (1996b), Norris & Ortega (2000), Bonk & Ockey (2003) and Brown (2003), among others. Bonk & Ockey (2003) took facets such as examinee, prompt, rater and rating criteria and analyzed these using FACETS many-facet Rasch analysis to check the extent to which they contributed to score variances and for fitness of the test performance model. Their findings suggested that the controlled group oral may be a viable alternative to estimate students' L2 oral ability when the alternatives such as the oral interview or multiple-choice test, especially when done on a large scale, are more expensive and time consuming. Using discourse analysis on two interviews involving the same candidate with two different interviewers, Brown (2003) illustrated how variation in interviewer behaviour (in how they structure sequences of topical talk, their questioning techniques, and type of feedback provided) affects

the candidate's perception of the task, and ultimately their response. More importantly, because communication between interviewer and candidate is co-constructed, effective communication can only be achieved if this is realized in the interviewer training for such interaction-based tests, and the test design itself which should spell clearly and unambiguously conditions and operations relevant to the construct. This is echoed in Swain (2001) who made the point that in a group, performance is jointly constructed, in terms of cognitive and strategic processes, and distributed across the participants. He asserted that information regarding this type of performance is invaluable in validating inferences drawn from the test scores. Other works on the problems of co-construction of discourse that can arise in group interactive tasks and tests include McNamara (1997), Fulcher (2003), Luoma (2004), and Dimitrova-Galaczi (2004).

All these studies point to the fact that the group oral test has similar concerns of the interview test such as rater reliability and practicality, and other concerns such as affective factors and the co-construction of discourse. It is also recognized as having great potential for assessing the oral ability of language speakers, and in the present research, one such test is investigated for its test worthiness in terms of context, theory-based and scoring validity.

Hence, largely in response to the shortcomings of the oral interview test, in particular the practical problems of test administration, and also the reliability of scoring, and more recent concerns with the group oral test, a move to a semi-direct form of testing speaking began.

2.3.2 The semi-direct speaking test

As stated earlier, the OPI faced some practical criticisms in terms of reliability in its rating process, the validity of its construct measurement and practical limitations. In order to address these problems, the oral test referred to as 'semi-direct' by Clark, 1979 (in Ellerton 1997), was developed in which the candidate is required to respond verbally to recorded stimuli, and his recorded responses are kept for subsequent scoring by trained assessors.

Historically, the idea of a 'semi-oral' test was suggested by Roach (1945) in his report on the oral examination that had been included in the Cambridge Certificate examinations. However, Roach was proposing a test where the examiner speaks and the candidate writes down his response, what is now normally known as a form of listening comprehension test (Spolsky 1990). Much later, in the UK, the Association of Recognised English Language Schools oral test (ARELS Examination Trust) was developed in 1966 (in Ellerton 1997), and the Test of English for Educational Purposes/TEEP was developed in 1983 (in Weir 1988); all of these were tape-mediated oral tests. In the USA, the tape-mediated Test in Spoken English/TSE was developed by Clark & Swinton in 1979, the Recorded Oral Proficiency Examination/ROPE by Lowe & Clifford in 1980, the Simulated Oral Proficiency Interview/SOPI by the Centre for Applied Linguistics/ CAL, 1982, and the Speaking Proficiency English Assessment Test/SPEAK by ETS in 1985.

The generic term SOPI (Simulated Oral Proficiency Interview) came from the Centre for Applied Linguistics/CAL where it was developed. Their first SOPI test was the

Chinese Speaking Test (Clark & Li, 1986), and others were later developed for Portuguese, Indonesian, Hebrew and Hausa (Stanfield 1989).

Claims of the semi-direct test

One of the main reasons why the SOPI was developed was because of the logistical difficulty the OPI poses for both large-scale testing and in testing the less commonly taught languages. The SOPI has major advantages over the OPI. For example, while the context of the OPI differs somewhat depending on the interviewer and examinee, the SOPI, by contrast, is uniform. All candidates receive the same input and instructions from the tape recorder and test booklets; this addresses, in part, some criticisms by Bachman (1988) and others, of the constantly changing nature of the OPI. The SOPI also addresses other difficulties of the OPI such as finding unbiased interviewers, practicality, and affordability. Hence, the SOPI, which was based on the original OPI test, became popular among test users especially because of its practical advantage of testing hundreds of candidates at a time, and its ability to reduce rater bias. The reliability of the SOPI has been its major advantage. Research on it has also focused on comparisons with the OPI and has yielded high correlation coefficients, such as by Clark & Clifford (1988), Stansfield (1989, 1991), Wigglesworth & O'Loughlin (1993), Kuo and Jiang (1997), among others, indicating a high degree of concurrent validity for the SOPI. These studies demonstrated that where it is not practicable to administer the OPI, a taped test could be substituted. SOPI lends itself to double marking and in the case of disagreements, there is a permanent record left for recourse to a third party.

The SOPI has also been used in a variety of contexts, such as for certification of bilingual education teachers in Texas (TOPT) (Stansfield & Kenyon 1991), and as

proficiency tests for the three most commonly taught languages, Spanish, French and German.

While it was thought that the semi-direct test could solve some of the problems of the OPI, unfortunately this approach to spoken language testing was not without its problems too. A significant theory-based issue raised about the semi-direct speaking test is the lack of reciprocity on the part of the examinee in the interaction. His/her role is largely confined to responding to questions or situations with little or no responsibility for participation in the interaction, i.e. initiation, continuance, topic shifting or termination.

In the case of the SOPI, surprisingly little research has addressed its construct inadequacy or coverage. Few researchers have criticized it (see Malone 1999), but because it often mirrors the techniques to a certain extent of the OPI, the criticisms of the OPI's lack of validity are also applicable to the SOPI. For example, like some direct tests, the SOPI's validity must be in question because of misgivings regarding the validity of the ACTFL guidelines used for rating; the construct validity of the test is affected by the validity of its rating instrument. Hence, criticisms of the SOPI have often arisen from studies that compare the OPI and the SOPI in concurrent validation studies by Clark (1979); Lowe and Clifford (1980); Shohamy and Stansfield (1990); Stansfield and Kenyon (1988, 1989, 1992b) and Wigglesworth and O'Loughlin (1990).

While correlation coefficients between these tests are usually high and concurrent validity is inferred from this, the question of the theory-based validity of the SOPI remains. High correlations between the two tests may indicate the two approaches

are saying similar things about the candidates, but this is not sufficient evidence that they measure speaking ability. Performance on an indirect test does not necessarily have the theory-based validity a more direct test might exhibit.

Secondly, the content validity or coverage of the SOPI is questionable, as is the case with the OPI where queries have been raised as to whether a single type of interaction (i.e. an interview) is sufficient to assess oral proficiency (Shohamy 1983; Perrett 1987; Raffaldini 1988; van Lier 1989). While the SOPI consists of a series of set tasks which elicit oral discourse through aural and visual stimuli and the scoring is done retrospectively, the problem of making inferences about test scores from a semi-direct test to the real life oral ability of a candidate remains unanswered. Despite potentially wider content coverage permitted by semi direct tests in terms of skills, it must be the case that performance conditions will differ markedly between direct and semi direct tests. For example, there will be differences in the effects of the response format where in SOPI, candidates construct a response to recorded stimulus and in OPI, candidates respond to questions or statements made by the interviewer; the effects of these differing response formats on processing can potentially affect test scores. Another difference is variability in time constraint, especially in the OPI, and in SOPI where limited time for each task can affect performance and candidates internal processing of his executive resources. The interlocutor is a major variable that affects performance as candidates monitor their own performance based on input from the interlocutor; factors such as gender, speech rate, and accent affect the performance of candidates in the OPI, but in the SOPI, this does not seem to be a problem. These differences point to the fact that performance conditions vary widely between the OPI and the SOPI, and this has validity implications.

More recently, Lumley & O'Sullivan (2005) investigated the impact on SOPI performance of systematically manipulating tasks (speaker + topic bias) and found little evidence of meaningful effects. Hence, as stated above, when performance conditions are changed or manipulated, performance may or may not be affected; in the case of the SOPI, the interlocutor effect appears to be small.

Validity is concerned with the statements that we make about a candidate's ability from the test results that we obtain; how test tasks that are limited to aural and visual stimuli enables a candidate to demonstrate his ability of language use which reflects the domain of target language use, needs to be investigated further. How a semi-direct test is able to capture features of oral proficiency other than pronunciation, tone, grammar, and structure, is also open to investigation. In other words, what is the theoretical basis for making a decision of a candidate's oral ability based on a recording of his voice over several tasks?

Even though no comprehensive validation of these tests ever seems to have taken place, there is sufficient criticism in the literature to raise concern about the SOPI as well as the OPI. The OPI and SOPI would not seem to fully meet the criteria of content validity, theory based validity, reliability and consequential validity (see Weir 2005). Despite the absence of a suitable framework for validating a speaking test from the disparate studies we have looked at, it already seems clear that neither OPI nor SOPI are, in their present form, capable of providing a fully satisfactory approach to testing spoken language.

What seems to be necessary is an approach to testing spoken languages that is content valid, reflects real life processing in speaking, has high reliability, matches other estimates of speaking we have some confidence in, and also has a good washback effect on individuals and society.

A potentially interesting recent development in the field of testing is the use of computer technology for the administration and scoring of tests. This development at first sight seems to have the potential to capture the validity of the direct OPI test and the reliability and practicality of the semi direct SOPI. Literature in the field of computer-based or web-based testing is, given its novelty, not surprisingly limited, and work on developing the test of oral proficiency using computer or web-based technology is almost non existent.

The concerns that language testers have with computerised testing (CT) must of course correspond to those discussed in relation to other forms of tests like the direct or semi-direct test; whether computerized technology can be used to deliver an oral proficiency test that is reliable, valid and practical, is the major concern.

2.3.3 Computerized testing

Generally, tests that are administered at computer terminals, or on personal computers, are known as computerized tests. The history of computerized testing, began as far back as the 1930s when the first IBM model 805 was used for scoring objective tests, specifically those comprised of multiple-choice items (Fulcher 2000:93). This ability for automatic marking and the major concern for assessing large numbers of people cheaply and efficiently spurred many specialists of testing and assessment to use the computer. While these early uses of computers

concentrated on data management and analysis only, developments by the 1970s led to extension of the use of computers to delivery (Fulcher 2000: 94). Since that time there have been two main modes of delivery via computer. The first of these is Computer-based Test (CBT) where the computer mimics the traditional test and serves merely as a better means of delivery of a single common test (Chalhoub-Deville 1999; Fulcher 2000; Roever 2001; Chalhoub Deville 2001). Then in the 1980s, the first Computer Adaptive Test (CAT) was implemented by the College Board Graduate Record Examination. In CATs, the computer plays a more sophisticated role through item selection and tailors each test to the ability of the candidate taking it (Dunkel 1999; Fulcher 2000;Norris 2001). Both of these developments are described in detail below.

2.3.3.1 Computer-based tests

A Computer-based test (CBT) is simply a paper-based test that has been put on a computer in the same linear fashion as the paper-based version. They are delivered on individual computers or closed networks and as a testing medium, they have significant advantages. CBTs can be offered at any time unlike mass paper-and-pencil administrations which are constrained by logistical considerations. In addition, CBTs consisting of dichotomously-scored items can provide feedback on the test results immediately upon completion of the test, as well as on each test taker's response. This feedback could have a positive impact on the curriculum as certain items can be 'tagged' to provide test takers with feedback on particular areas of strength or weakness (Davidson 2003). The integration of media enhances the testing process itself, and enables the tracing of a test taker's every move; this provides examiners with valuable information about testing processes as part of

overall test validation (Roever 2001). Examples of language tests that are computer-based are CBT TOEFL, the Graduate Record Examination/GRE, and the Scholastic Aptitude Test/SAT.

The main advantages that a CBT has over other types of technology-mediated oral proficiency assessment like the SOPI, are in the examinees' affective responses, and efficiency in administration and scoring (Brown 1997; Dunkel 1999; Norris 2001). The CBT is distinctly advantageous over the SOPI, or the direct test OPI, in terms of eliciting examinee performances; the computer does away with the need for a test proctor or interviewer to distribute and collect test materials, to monitor and conduct test activities, to capture examinee performances, and so forth. The most useful technological advance featured in a CBT is the replacement of tape recordings with computer-based digital audio or video recordings. This will benefit raters tremendously, as they will be able to listen or watch examinee performances on test tasks in any order, in part or whole, with instantaneous repletion of particular segments of speech, etc., all without the cumbersome demands of forwarding or rewinding tape recordings used in SOPIs and OPIs (see Norris 2001).

2.3.3.2 Computer adaptive tests

CAT is considered the most important development of the last decade in interactive testing. In a CAT, the computer branches to certain sub-tests or selects the next test item, depending upon the response pattern of the individual test taker. This is made possible by the extensive use of Item Response Theory, and the development of algorithms that drive the programme to select and deliver test items, score responses, and provide immediate feedback to test takers (Dunkel 1991,1999). Its

main advantage is in the efficiency of the system which enables examinees to complete only about one third of the items, compared to the paper and pencil test. The only example of a Computer Adaptive (CAT) speaking test is the Computerized Oral Proficiency Instrument (COPI) (see Kenyon & Malabonga 1999). It is a multimedia, computer-administered adaptation of the SOPI which is audio-delivered, and like the SOPI, simulates real life tasks to elicit speech to be rated using the ACTFL Guidelines' criteria. Because the COPI is an adaptive test, it has a pool of approximately 100 assessment tasks that the candidate can select from; each task has a targeted ACTFL level (Novice, Intermediate, Advanced or Superior) and is coded for its function and topic/content area. It is also clear from the description that developing the CAT is a highly technical and huge task, and the actual test takes a long time (anywhere from 30-50 minutes per candidate). (from Kenyon & Malabonga 1999 on the rationale and operation of the COPI)

2.3.3.3 Tests on the web

These are computer-based tests, which are delivered via the World Wide Web (www). There are two main types of test that could potentially be delivered on the web. Firstly, a Web Based Test (WBT) can be written in files using HTML, stored in the server and downloaded to the test-taker's computer, all at once or item-by-item. These are essentially low-tech tests in that they run completely client side and use the server only for retrieving items and storing responses. Post test feedback is available only where items are objectively scored (Roever 2001).

The other type is the high-tech Web Adaptive Test (WAT). These tests make heavy use of the server to handle item selection through adaptive algorithms and to collect

and analyze test taker responses (Roever 2001: 85). Unfortunately adaptive tests, both CAT and WAT, require huge item or task banks such as that being developed by ETS Princeton for tests such as TOEFL, over a period of ten years. Recent experience in China has also demonstrated that even a 10,000 item bank may be insufficient to guarantee test security (Luecht 2001). Thus WATs are clearly impractical even for large scale testing organizations, let alone the lone researcher. The only viable option is therefore either a computer-based or a web- based delivery system.

Given the context for the test in Malaysia, a web-based speaking test is the preferred option. WBTs are advantageous over traditional CBTs with regard to practicality and logistics in terms of

- flexibility as regards time and place of administration,
- delivery; in that they require only a free, standard browser for their display,
- cost ; a WBT is relatively inexpensive for testers and the test takers
- efficiency ; electronic script management and marking

Although there are outstanding issues that need attention, such as WBT validation procedures and the as yet unrealized potential of oral testing over the web, it is clear that the Web greatly expands the availability of computer-based testing with all its advantages over other more traditional forms of testing. Examples of recent WATs include the DIALANG project, several in-progress reports on the Computerized Oral Proficiency Instrument/ COPI (Kenyon & Malabonga 1999, 2000, 2001, 2002), and on-going research projects at UCLA (Bachman et al, 2002).

There are other WBTs focusing on receptive skills (Jones, 2001) which range from Web-based placement exams, to Web-based language tests, exercises, and quizzes through easy-to-use authoring tools for language teachers and language learners. However, to date, limited work and effort have been put into the testing of oral ability in the field of computerized testing.

Thus, there is a distinct possibility that we can improve on the SOPI version of the speaking test through CBTs using web-based technology. To achieve this we need to take out the best from existing direct tests in terms of content, theory-based validity, and add the positive elements of web-based testing in terms of reliability and practicality.

In summary, the literature review above illustrated the following differences regarding the three methods of testing speaking:

a) The direct speaking test such as the OPI poses serious shortcomings in terms of its claims of the nature of the interview (context), rater reliability (scoring) and practicality

b) The semi-direct test such as the SOPI faces issues of nature and number of task (context) and cognitive processing (theory-based)

c) The computerized test, with advantages of new technology, could enable us to draw the best of direct test in terms of theory-based and context validity, and semi-direct test in terms of reliability and practicality to develop a speaking test that may meet the necessary requirements of validity, reliability and efficiency.

The proposed study on a computer-based speaking test will therefore explore the possibilities of delivering an improved version of the existing direct test using

computer technology. In essence, as stated in Chapter 1, given the scope, reliability and practicality, and validity issues faced by the existing speaking test at the university, an attempt to introduce the indirect test in Malaysia is justifiable, especially because it attends to the concern of standardization in large scale testing. Details of the study and its possible impact are discussed in detail in Chapter 3: Methodology.

As we also saw in the literature review, both SOPI and OPI have come in for a lot of criticism from many different angles (see Alderson et al. 1987, for examples). In order to even attempt a computer-based speaking test, we would need to investigate the existing direct test in terms of its validity aspects. Unfortunately, a systematic, comprehensive framework for validation has to date been absent from the testing literature. To meet this shortcoming this study will refine and operationalise a new socio-cognitive framework for the test of speaking (O'Sullivan 2000; O'Sullivan & Weir 2002) to validate the existing test, and to validate the new test to be administered via the computer. The framework takes into consideration the important elements of context validity, theory-based validity, scoring validity, consequential validity, and criterion-related validity. (this will be explained in detail later in the chapter).

This next section discusses issues relating to the concept of validity in language testing, leading to the final section on the socio-cognitive framework for validating a speaking test and its application in the study.

2.5 Issues of Test Validity

This section highlights the importance of what validity means as this is one of the main directions of the present study; the topic had not been addressed before by the university test developers and no systematic validation had been conducted on the tests under discussion. Hence, this is a first attempt at describing characteristics of the speaking test as it is being tested currently, and in doing so defining parameters of the test that are valid for its purpose.

A. Defining validity

In establishing the definition of test validity and the process involved in test validation, experts have gone from tagging labels to categorizing types of validity, taking a historical approach, establishing test score interpretations and determining consequences of testing.

Cumming (1995) for example, calls validity an 'ominous' word and test validation in language assessment is ominously important, but establishing the validity in language assessment is by all accounts 'problematic, conceptually challenging, and difficult to achieve'. Geisinger (1992) refers to changes in validity theory as a "metamorphosis", and Shepard (1993) refers to them as an "evolution". Shepard draws upon several earlier works to make her case for an argument-based approach to establishing test validity. These include Cronbach's (1971) original idea that validity resides not in the test but in the interpretation of the data arising from the test and the conception of validation as an evaluation argument (Cronbach 1988, 1989), Messick's (1989) model for a unified but faceted validity framework, and

Kane's (1992) conceptualization of validation as the evaluation of interpretive argument.

However, Borsboom et al (2004) argued strongly that validity literature of the past fifty years has not been about the simple, factual question of whether a test measures an attribute. It has been about the complex question of whether test score interpretations are consistent with theories and observations (beginning with Cronbach & Meehl 1955), or with even more complicated system of theoretical rationales, empirical data and consequences of testing (Messick 1989). They offer a simpler conception of validity: that a test is valid for measuring an attribute if the attribute exists and if variations in the attribute causally produce variation in the measurement outcome; not complex, faceted or dependent on nomological networks and social consequences (Borsboom et al, 2004: 1061). Whichever the case, the question of whether a test measures what it really intends to measure has been debated for a long time amongst experts in educational measurement and language testing fields.

It seemed that only in the past two decades has validity been discussed in the literature in clearer terms so evidences of validity in a test can be empirically gathered and systematically produced (Anastasi 1988; Bachman 1990; Bachman & Palmer 1996; Alderson 93/95; Weir 1993, 2003/05; Hughes 2003).

From the historical perspective, Spolsky (1990, 1995) cites the Roach report (J.O. Roach 1945) as the best treatments in print of the way that non-psychometric examiners attempted to ensure fairness in subjective traditional exams (oral or written). The report called for better standardization and coordination of the exams,

and a clearer description of the standards used for rating; this was the key issue in terms of test validity at the time. Similarly, Weir (2003) describes how the Cambridge Certificate of Proficiency in English (CPE) has been in the testing field since 1913. While it also had an oral test which comprised reading aloud and conversation, little attention was paid to the nature of the speaking component until several decades later in 1975, 1984, and 2002 when the test was further revised (ref Weir & Milanovic (eds.) 2003 for full history of the development of the CPE).

B. Types of validity

Since the 1950's (Cronbach & Meehl) then, validation in educational measurement had been concerned with establishing procedures which ascertain the justification for inferences and uses of tests. Test validity was viewed as the extent to which test results/scores are an accurate representation of the inferences that we make about a candidate's true level of language knowledge or skills (Messick 1980, 1989; Anastasi 1986, 1988; Bachman 1990, 1996; Grondlund & Linn 1989). Kane (2001) discussed how the concept of validity had evolved from the early days of a criterion-based model (Cureton 1950; Cronbach & Gleser 1965). This then moved to more content-based models (Ebel 1961; Angoff 1988; Messick 1989), and the construct model of validity which was the period 1955-1989 when the emphasis was on construct validity as a unified framework or model for validation (Messick 1975, 1988, 1989; Cronbach 1988). For many years however, this idea of construct validity as the centre for a unified model had its drawbacks and difficulties (see Cronbach 1988 for further details on the distinction between 'weak' and 'strong' programme of construct validity). Subsequently, these earlier principles fit naturally into an argument-based approach to validation (Cronbach 1988; Kane 1992; Shepard 1993). This was the approach later adopted by TOEFL (ETS research March '04) in validating the new

generation TOEFL tests; largely based on Kane's (1992, 2001, 2002) and Mislevy's (2003) approaches to development of propositions for validity argument based on a framework for potential inferences that might be made from test scores, as opposed to requirements of the Standards 1999 which was seen as a "smorgasbord" of options for approaching validation.

The philosophical changes in validity over the past decades have not only come from various sources, but mainly consist of categorizing and describing validity without propositions on how test validation is conducted. The 1954 test standards (APA 1954) were the first in fact to make the proposition when it listed four types of validity: predictive, concurrent, content and construct (later merged to three); the type of validity chosen for test validation is dependent on the test purpose (p737). Further revisions were made to include different types of validity standards that educators and researchers might use for their tests (APA/AERA/NCME 1966, 1974, and 1985). There was a shift in the literature on validity and test validation: validation was a strategy based on test use/purpose and consist mainly of four types of validity, which then moved to validation as a strategy based on inferences made regarding test scores and all types of validity as a unitary concept. In addition, there was a strong push for the integration of validity concepts under the umbrella of construct validity; construct validity was seen as central to the notion of validity (Messick 1980, 1981; Cronbach 1988; APA/AERA/NCME 1985 test standards). Still, while the 'whats' and 'whys' of changes in validity had been discussed, the issue of how these changes can be applied to validation practices had been given minimal attention.

Reports have shown the extent to which testing practices had actually complied with the standards of 1985. Qualls & Moss (1996) showed how documentation of reliability and validity evidence was often lacking in reports for scores on test instruments, and that such reports are infrequent and scarce. Johnson & Plake (1998) conducted an in-depth study to determine if changes made in validity theory (translated in the form of test standards) actually affected test validity practices. They reported that test standards do seem to be influential in forming measurement professionals' overall concept of validity, but are not as influential in determining the actual validity requirements that should be applied (p 751).

What seemed clear was that the future impact of test validity on validity practices lay in determining what test validity studies should consider when providing validity evidence, not just in determining the steps that were required for validity studies.

C. Validating the construct

In more recent years, validity is seen as being multi-faceted in that the varieties of evidence supporting it are not distinct entities; they are complementary to each other and all together constitute the evidence needed to ensure the validity of a test (Messick 1989, 1995). Hence, the term validity is itself used as the superordinate to encompass the elements: content validity, theory-based validity, marker reliability, consequential validity and criterion-related validity, and what evidence they can generate in support of interpretation of the test scores produced (Messick 1996, Weir 2004/2005).

Test validation then, requires more than just claims that test designers or writers make of the test. What we need is evidence from a variety of sources, both quantitative and qualitative in nature, gathered at different stages of testing. At the stages of test design and development, the test should be constructed to an explicit specification, which addresses both the cognitive and linguistic abilities of language use, and the context in which these abilities are to be performed (theory-based validity and content validity). At the implementation stage, we need to apply statistical analyses to the data produced by the test to show the reliability of the test, and finally, after the test, we can collect more data to ensure that inferences we have made about underlying abilities based on the test scores are justifiable (Weir 2005). Thus, we need to conduct validation processes both before (a priori), with emphasis on content and theory-based validity, and after (a posteriori) the test is administered, with emphasis on marker reliability, consequential validity, and criterion-related validity. Only when all these measures have been taken that our test has evidence-based validity, i.e. the relationships between the test instrument and the construct(s) it attempts to measure.

In the past decade, researchers in the field of language testing have conducted investigations into establishing the validity of tests, whether small or large-scale in nature. Among them, include investigations into the validity of the OPI (Bachman & Palmer 1981; Dandonoli & Henning 1990; Shohamy 1994), validity of other speaking tests such as the TOEFL Test of Spoken English/TSE (Powers et al 1999), the Norwegian EVA speaking test (Hasselgren 2002), the Simulated OPI (Kenyon 1998), dialogue within small groups (Swain 2001), the MATHSPEAK test (Douglas & Selinker 1993). Most recently, Weir (2005) highlights how the Common European Framework of Reference (CEFR) lacks clear descriptions in terms of the demands of

theory-based validity, context-based validity and scoring validity for each of its six CEFR scale levels.

To borrow the definition from (Thomson 2001): "Validity is the extent to which the interpretation and use of an assessment outcome can be supported by evidence". It is what provides the assessments with quality and is concerned with the soundness of the interpretations and uses you make of your assessment results, not a property of the assessment instruments themselves.

D. Building a validity argument

As stated earlier in this, and the previous chapter, the socio-cognitive framework for validating the speaking test (Weir 2005) which constitutes five components of validity, (described below), is used in the present research from the initial stage of research instrument development (the questionnaire & interview framework) to the actual process of establishing validity elements of the test (data gathering). Such a framework is essential and important if the study is to establish whether a test is 'valid' as a testing instrument and for its specific purpose(s), i.e. it has been developed, conducted, analysed, and used again, according to very clear specifications based on professional judgements and sound theoretical background.

2.4 The Socio-cognitive framework for validating skills in language tests

2.4.1 A model/framework for validating a speaking test

The socio-cognitive framework for validating the speaking test (Weir 2004/2005, see diagram below) was used as the major reference by which the research instruments of the study were developed. It was refined and operationalised for use at the design

and developmental stages of the study, through the implementation and post-exam stages of the direct and computerized speaking tests, i.e. the findings were reported according to validity elements of the framework. As stated in chapter 1 and repeatedly in the current chapter, a framework for validation is needed in order for the researcher to be able to gather data systematically and objectively. The framework was utilized as it is by far the most comprehensive framework of its kind in the literature; hence, it was operationalised in the development of the speaking test questionnaires and interview guidelines used for the study (see Chapter 4 Methodology on 'direct test validation' for details on how the instruments were developed using the framework). Work on developing the framework began in early 2000 (see Weir & O'Sullivan 2002) and this continued at the Centre for Research in Testing, Evaluation and Curriculum in ELT (CRTEC) of Roehampton University, during which focus group discussions were conducted with other associates of the centre. These instruments were used as the main sources of data gathering for the study, and each of them contained items which encompass each validity component and elements found in the framework. However, as noted in Chapter 1 earlier, in the present study, data was gathered mainly for context, theory-based and scoring validity. This was decided upon for two reasons: it was beyond the scope of the study to gather data on all components due to time and logistical constraints; though some data were initially gathered on consequential and criterion-related validity, they were not sufficient to merit as contributory factors of test validity. Secondly, it had been reiterated that in test validation, construct validity refers to the test task & its features (context validity), the test taker & his/her strategies for attempting the task (theory-based validity) and factors associated with how the test is assessed (scoring validity).

The framework consists of five types of validity evidence required for the above purposes: context validity, theory-based validity, scoring validity, consequential validity, criterion-related validity.

Figure 2.2 A Socio-cognitive framework for validating speaking tests (Weir 2005)

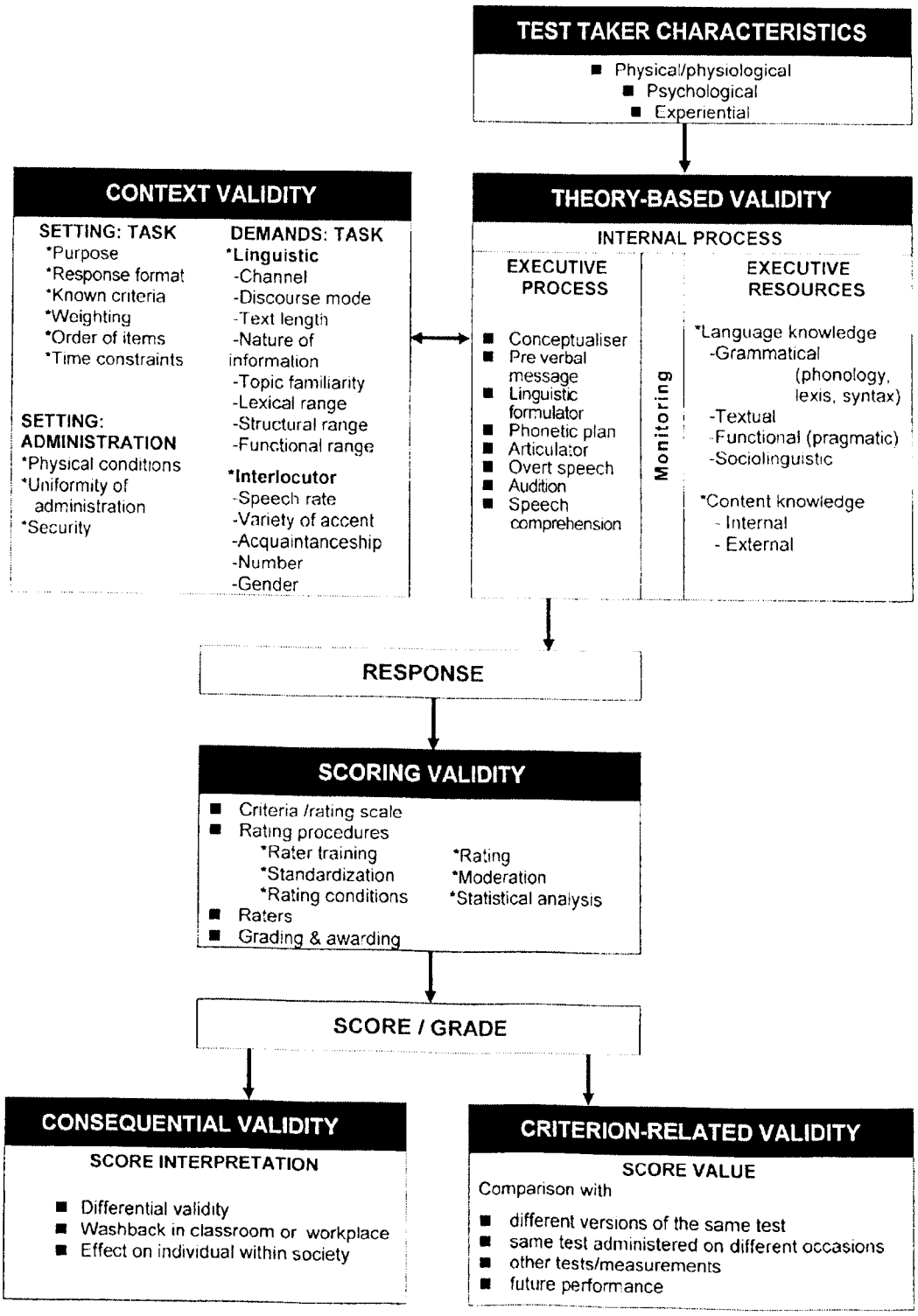


Figure 2.2 above shows the conceptual framework and its various components, and an explanation of each component follows.

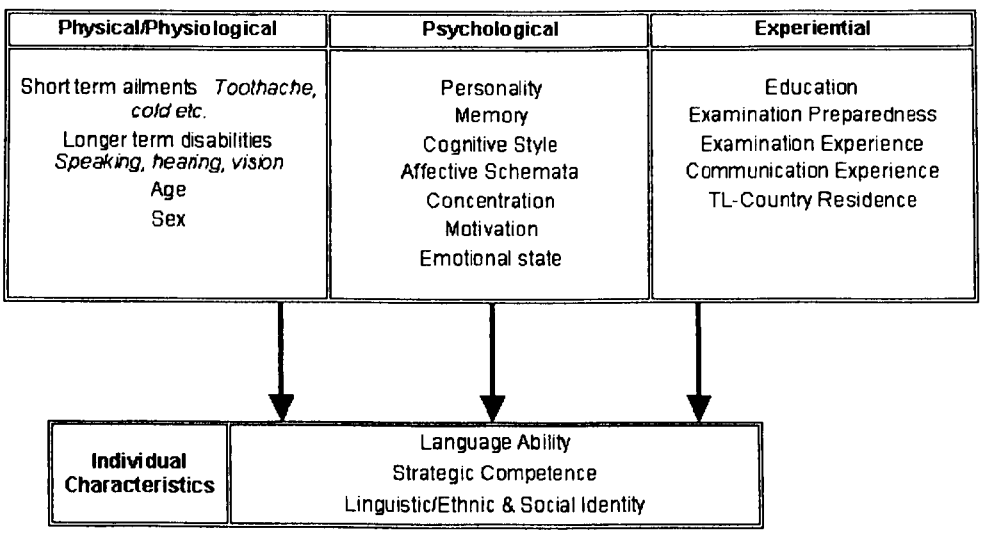
2.4.2 *Components of the framework for validating the speaking test*

As mentioned above, at the stage of design and development of the test, we are concerned with evidence for theory-based validity and content validity. We must, however, consider the test taker who is central to the validation process first, as the range of factors associated with the test-taker may have potential effect on test performance. O'Sullivan (2000a), drawing on earlier frameworks (ref Bachman1990; Cohen 1994; Alderson, Clapham & Wall 1995; Bachman & Palmer 1996; Brown 1996) suggests an alternative view of test taker characteristics. This alternative view is summarised in the table below.

Figure 2.3 Summary of alternative framework suggested by O'Sullivan (2000a)

O'Sullivan (2000a)	When identifying a model representing factors which affect performance suggests that Characteristics of the Test Taker (CTT) might be described under the headings: • Physical/Physiological • Psychological • Experiential

Figure 2.4 Characteristics of the Test-Taker (based on O'Sullivan 2000a)



a) Theory-based validity is related to the internal mental processes and involves executive resources and executive process. Executive resources include linguistic knowledge/competence and content knowledge of the speaker. The speaker attends to the task using his/her grammatical, discoursal, functional, and sociolinguistic knowledge of the language. These are equivalent to Bachman's (1990) language competence components of organizational and pragmatic competence. He defines language ability as having two components, language knowledge/competence and strategic competence or metacognitive strategies, which combine to provide language users with the ability to create and interpret discourse in a task or test. In addition, the speaker brings to the test event, content knowledge, which relates to background/prior knowledge of the topic or content, and knowledge he gets from the content of the test itself.

In the speaking test, internal processing for the speaker is complex and dependent on the input he receives. Drawing on Levelt's (1989, 1993; Levelt et al. 1998, 1999) blueprint for the speaker, the theory of speaking presents processing components and their processing

systems where each component is specialized and autonomous to allow for the generation of uninterrupted fluent speech.

In a speaking test where the interlocutor or examiner has control over the input, as well as the format of the test, the speaker has to attend to this input by accessing his/her executive resources, and internally processing this knowledge and information, to produce speech which is appropriate within the time and context given for the task.

This complex process has led some researchers to investigate further the internal processes that a speaker undertakes in speech production during a speaking test. Studies using various introspective methods have been conducted as means to elicit data about thought processes involved in carrying out a task or activity by Dornyei & Kormos 1998, Cohen & Olshtain 1993, Mackey, Gass & McDonough 2000, Brown 1993, Cohen, Weaver & Li 1995, Poulisse 1990, and Swain & Lapkin 1998. In general, they confirm the view of speaking proposed by Levelt, but also revealed a number of additional strategies across tasks.

The relationship between theory-based validity and content validity is a symbiotic one; the context in which the test task (input) is presented will influence the internal process of the speaker. For example, the mode of input, whether it is listening to a recorded speech or looking at diagrams or pictures, will affect how the speaker conceptualizes and processes these messages as pre-verbal message.

b) Context validity is concerned with context coverage, relevance and representativeness. According to McNamara (2000), it is "the extent to which the test appropriately samples from the domain of knowledge or skills relevant to performance in the criterion." Context

validity for the speaking test is divided into: the characteristics of the task, and setting or test administration. Task is further divided into rubrics, discoursal variables, linguistics variables, and interlocutor variables, and test administration is divided into physical conditions, security, and uniformity of administration. The requirement for the test development team is to present as much evidence as possible for each of these elements, i.e. that test items or tasks chosen are based on or have fulfilled each of the criterion in this framework. For example, if it can be shown that the response format such as a short presentation followed by a discussion, reflects the instructional objectives of a course on speaking for academic purpose, then this is one piece of evidence for the choice. In addition, and more importantly, whether tasks reflect the domain of language tasks necessary for academic language use needs to be established. The interlocutor variables are crucial in a speaking test, especially in the direct test where interaction takes place between the participant and an interlocutor, who may, at times, be the examiner. Factors such as interlocutor's speech rate, gender, number, and acquaintanceship may have an effect on performance. Physical conditions, security, and uniformity of test administration must be ensured for the test to be fair for all participants, and valid.

Theory-based and context validity are connected to each other; the input from the test task will impact on the cognitive process of the speaker, in varying degrees, while having to draw from his/her internal and external resources for language and content knowledge. It is crucial that test content is relevant and representative of the skills or abilities that we would like the speaker to demonstrate within the test context.

c) Scoring validity is one aspect of validity evidence that needs to be established at the implementation stage of the test. It is an unavoidable assumption that a candidate's observed score is not entirely due to his language ability but is also influenced by factors

other than the language trait in which we may be interested, i.e. sources of error, or unreliability. It is the identification and minimization of such errors of measurement, actual or potential, which must concern the test developers (Ellerton 1997). The literature shows the importance of considering the reliability of scores produced by a test in Cronbach (1984), Bachman (1990), Weir (1993), Alderson et al. (1995); Bachman & Palmer (1996), Chapelle (1996), McNamara (1996), and in AERA, APA & NCME (1999). Rating is an important factor affecting the reliability of a test; it involves the criteria/rating scale, rating procedure, raters, and grading and awarding.

Rating is a special concern in the testing of speaking in situations involving the direct test, such as an interview and discussion, where rating is conducted as the task is being performed. In this situation, the major concerns are rater reliability, and how marks are awarded based on a rating scale or instrument. Hence, the rating procedure (Upsher and Turner 1999; McNamara 2000), criteria/rating scale (Chalhoub Deville 1995; Fulcher 1996;North & Schneider 1998; ALTE 1998), rater training, rating conditions, moderation, and statistical analysis, are conditions that need attention in order to reduce potential errors in measurement. In semi-direct tests, this source of error in measurement is potentially reduced as the task is usually recorded, on tape or video, and rated by trained assessors later.

d) Consequential validity is related to the effects or impact of a test on various factors such as groups of test takers, teaching and learning in the classroom, and on society as a whole. The effect of a test, also known as washback has been studied by researchers such as Bachman 1990, Bailey 1996, McNamara 2000) and it refers to the effect or impact of a test on classroom teaching and learning. It could be a positive washback if the effect is positive and encouraging, or a negative washback if it is not.

A test can affect behaviours of other parties outside of the classroom, such as the departments of education, institutions of higher learning, parents, and other stakeholders who may have influence to change educational policies, based on the outcome of a test. Differential validity refers to the effect of differential item functioning that "exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test" (AERA, APA & NCME 1999). The aim is to ensure that test bias does not occur.

e) Criterion-related validity is an area where validity and reliability overlap and refers to how the test relates to other external measures of the same ability (as described in Hughes 1988, Anastasi 1988, Messick 1989, Bachman 1990, Weir 1990, 1993). Criterion-related evidence tends to be divided into predictive validity, which refers to the degree to which we can predict from a previous test performance an examinee's future level on some criterion, and concurrent validity, which is the extent to which we can assess the examinee's current level on the criterion of interest. In both elements of validity the criterion may consist of scores from other tests, or candidates' self-assessments of their language abilities, or ratings of candidates by teachers, subject specialists, or other informants (Alderson et al. 1995).

Criterion-related validity is established if a relationship can be demonstrated between test scores obtained from the same or different versions of a test administered to the same candidates in the same conditions on two different settings (Weir 2005). This is done by comparison with different versions of the same test (alternate forms, equivalent forms, parallel forms, comparable forms), or comparison with the same test administered on different occasions. As shown in the literature review, studies on semi-direct tests have

often been conducted in comparison to the direct/live oral tests in various concurrent validation studies.

The final section in this chapter lists the research questions which were formulated based on the above review of the literature on the spoken discourse, testing of speaking, computerized testing and validity in language testing, and with focus on its theoretical and practical aims.

2.6 Research questions

Given the widely dispersed context in which the direct face-to-face speaking test is administered at UiTM Malaysia, the need for a more efficient and valid speaking test is pressing. This study will attempt to establish the validity and practicality of a semi direct computer-based speaking test by establishing a theoretical framework for validating the existing direct test, and comparing it with a new semi-direct web-based test. Thus, the major question that the study will be addressing is:

Can an operationalised framework for validating tests of speaking provide an evidential basis for replacing a direct test of speaking ability with a semi direct web based speaking test?

This is then divided into the following questions:

1. To what extent is a face-to-face speaking test valid in terms of:
    a) context validity
    b) theory-based validity
    c) scoring validity

2. To what extent is a proposed semi direct computer-based speaking test valid in terms of:
    a) content validity
    b) theory-based validity
    c) scoring validity

3. Can we justify replacing a direct test with a semi direct test of speaking ability?

As stated in Chapter 1, the proposed study will focus on the three components context validity, theory-based validity and scoring validity as the major components for validating the underlying construct, i.e. speaking. As Weir (1993) stresses, the conditions under which the test task is carried out (context validity), affects performance or how the task is accomplished by the test taker; more recently (Weir 2004/2005) how cognitive processing is affected by the context of the test and elements in scoring validity such as criteria and criterion levels that match elicitations demanded by the task.

The study is then conducted in line with the research questions formulated and derived from the literature on testing of spoken language in this chapter. The entire study is then organised and formulated accordingly in order to address these questions, using the framework as its backbone, from research design to instrument development and data collection and analysis.

The next chapter on Methodology will present details on how the study is conducted in terms of the various stages involved, and at each stage the participants, instruments and other resources that the study incorporates. These will be described in terms of how they were selected, developed and used.

Chapter 3: METHODOLOGY

3.1 INTRODUCTION

This chapter presents the methods used for conducting the study & the design that it follows. At the beginning of the present study on 'Testing spoken language using computer technology', a framework for validating language tests was developed and then used as the stimulus for developing research instruments and all other facets of the study. In addition, the study included two major validation exercises on validating the direct speaking test and the computer-based speaking test (**see Sections I - IV below** for details) which adopts the framework. Weir's (2005) framework for validating language tests was used in the study from the start and throughout the study. The instruments used for data collection were developed in accordance with components and elements of the framework, data collected were organised and analysed according to various validity components of the framework, and conclusions about test validity and variables that affect test performance are made with reference to the framework. The study was conducted in line with the research questions which were formulated based on the literature on testing of spoken language (see Chapter 2 on 'Literature review' for details). The entire study was then organised and formulated accordingly in order to address these questions, using the framework as its backbone, from research design to instrument development, data collection and analysis.

The validation exercise is a critical stage in the study because we need to identify the strengths and weaknesses of an existing test, and this would be the basis for the

development of a new semi-direct, computer-delivered test of speaking (see Ch I on 'Introduction' for background information & main purpose of the study).
Recent literature on research methodology calls for researchers to have a less rigid, one-track method and open to a wider use of methods (Gorard & Taylor 2004; Robson 2002; Cohen et al 2000 among others).

In the present research, the validation process essentially adopted a *triangulation method,* which is the use of multiple sources to gather data, and according to the literature on research methods, this enhances the rigour of a research (Robson 2002). More importantly, it is a strategy that can help the researcher to deal with possible threats to the validity of the research; threats such as inaccuracy or incompleteness of data, or imposing your own meaning to what is happening rather than what occurred or emerged during your involvement with the setting. The study included two types of triangulation (Robson 2002; p174): *data triangulation* which is the use of more than one method of data collection (interviews, use of questionnaire, observations, etc), and *methodological triangulation* which combines quantitative & qualitative approaches. Through triangulation, we can attempt to counter threats to the validity of our data collection process. It is noted too that this method can open up possibilities of discrepancies and disagreements among the different sources; interviews and documents may be contradictory, and two observers may disagree about what has happened.

In order to overcome such problems in the present study, the researcher had taken steps such as trialling & piloting questionnaire and interview items, interviews were recorded

and transcribed, and questionnaire data were counter-checked as a data entry assistant entered them into the SPSS programme. All quantitative & qualitative data were analysed and details are reported below. In essence, there was little conflicting information between various sources and this has contributed to a detailed examination of the features of the speaking test. The main sources of evidence for this validation exercise came from questionnaire survey, interviews, analysis of test documentations, observations, and test scores (see Appendix 3A in attached CD I).

From the start of the study, the validation process was planned, & the instruments were developed, according to the socio-cognitive framework for validating a speaking test (Weir 2005). The framework consists of five components of validity: context validity, theory-based validity, scoring validity, consequential validity, criterion-related validity (see Ch 2, section 3.1 on Socio-cognitive framework for validating language tests). As discussed in the chapter/literature on validity, the main purpose for validating a test is not just, so teachers and test developers can justify and have faith in the results that their tests produce. It is in fact, an attempt to ensure that the results of the performance on the test give us an accurate picture of the underlying abilities or constructs we wish to measure. It does not reside solely in the reliability of test scores, but it is now 'multifaceted', that is many types of evidence are needed to support any claims for the validity of scores on a test (Weir 2005: Ch 2). Validation thus begins from the time the test is constructed on a set of explicit specifications, which address both the cognitive and linguistic abilities involved when a candidate demonstrates them in a context or conditions specified by the test. Next, we need to look at the data generated after the test had been administered, apply statistical analyses to see how dependable the results

are, and finally we collect data on events after the test had been developed and administered to provide us with more information on whether the inferences we make on the basis of test results are justifiable. Hence, the validation process requires active participation on the part of the researcher to gather as much evidence possible in all aspects of the validity framework, using as many methods and sources possible to gain an in-depth, more detailed picture of the phenomenon under investigation.

For the present study, all validity components and their elements in the framework were included in the design of the questionnaire, and interview framework/guidelines. In addition, all other aspects of the process such as the reporting and the analyses of data collected, centred on the validity framework. This also made it a fixed design study in which from the early stages, a substantial amount of its specifications centred on the framework. However, as stated earlier, a substantial amount of data, especially in the first main study, had also been gathered through interviews, observations, and various focus group discussions, what is termed a more qualitative-type data collection process. This leads us back to the point of a multi-method or triangular approach of data collection.

This chapter will be presented according to the following sections as they provide a detailed account of the processes involved throughout the study.

SECTION I: Research design according to research questions

This section provides in detail the various stages that the study will undertake to address the research questions that were formulated at the end of Chapter 2. This can be called the 'theoretical' model as it was conceived of when the study first began.

SECTION II: Development of research instruments

This section provides details of the development of the research instruments used in the study, which include questionnaires, interview guidelines, observation and participation schedules; pilot studies, which were conducted at various stages of the study to try/test the instruments out, are included in this section.

SECTION 111: Description of participants, locations and research schedule

At the end of the description of participants and the locations involved in the study, a matrix is presented which shows the instruments used at each stage of the study.

SECTION 1V: Model of the research plan

Unlike the research design in section I, this model illustrates the actual conduct of the study from stage 1 to 5. This section illustrates and describes the data collection process and procedures that were carried out; it can be termed the 'operational model' of the study.

SECTION I: Research design according to research questions

As stated earlier, the study attempts to address the research questions that had been formulated from a review of the literature. At this point the following stages of the study were developed according to the research questions; these stages are described in detail below.

The main question that this study will address is:

Can an operationalised framework for validating tests of speaking provide an evidential basis for replacing a direct test of speaking ability with a semi direct web based speaking test?

This is broken down into three questions (see Chapter 2, section 2.6) which encompass the issues of validating the direct speaking test and semi-direct computer-based speaking test. Based on these questions, the study will comprise the following stages:

Stage 1. Develop a validation framework for speaking.

To do this, we need to:

1.1 Specify the elements of the construct of speaking in an academic and social setting.

For the specifications of the construct or the underlying models, we must consider theory-based validity and content validity largely through qualitative methods, which involve

- a literature review of the theories underlying the speaking construct
- a literature review on research on the testing of speaking

- content analysis of curriculum documentation (e.g. course syllabus, specifications, needs analysis)
- content analysis of existing tests of speaking
- qualitative feedback from students lecturers, examiners, administrators, and experts using a questionnaire.

1.2. Specify the components of scoring validity

For the specification of scoring validity, we have to consider the elements of rating through qualitative methods, which will involve

- a literature review on scoring validity
- qualitative feedback from students, examiners, lecturers, and expert judgement (on assessment criteria, rating procedures, raters, grading and awarding) using questionnaire and interview techniques.

1.3 Specify the components of consequential validity

To specify this aspect of validity, we need to investigate the components of score interpretation through qualitative methods that involve

- a literature review on consequential validity
- expert judgement and feedback from students, lecturers, administrators, on components of this aspect of validity evidence (washback in classroom, differential validity, effect on society) using questionnaire and interview techniques.

1.4 Specify the components of criterion-related validity

Specifying this aspect of validity evidence requires us to explore the methods of judging score value through

- a literature review on criterion-related validity

- feedback from administrators, lecturers and local authorities/experts from the examination board on judging score value, using the interview technique.

Stage 2. Develop and pilot instruments and procedures for applying the validation framework to the existing speaking test

In light of the theoretical framework established in Stage 1, we will develop and pilot instruments and procedures to validate the existing speaking test

2.1 Access evidence:

We need:

- documentation of special arrangements
- documentation of examinee test booklets

2.2 Administration evidence:

We need:

- checklist of physical conditions
- checklist and observation of uniformity of administration
- observation of security

2.3 Theory-based validity evidence

This evidence refers to the extent to which the hypothesized ability or trait is reflected by the measurement in the test. We need

- qualitative feedback from test takers through the use of a questionnaire and/or retrospective protocols based on videotaped sessions of the speaking test
- qualitative expert judgement of tasks (questionnaire: on the components in the speaking process)

- statistical analysis (multi-faceted Rasch analysis to investigate the impact of facets other than person ability and item difficulty on the rating process, and to check rater reliability)

2.4 Content validity evidence

Content validity refers to the extent to which the items or tasks represent the area of knowledge or ability to be tested. It reflects the representativeness of a test that is used to measure the ability or trait. We need

- feedback from students and expert judgement protocols (questionnaire on content coverage: rubrics, text type, type of information, topic, time, weighting, etc.)

2.5 Consequential validity evidence

This refers to "the practical consequences of the introduction of a test" (McNamara 2000). The instruments we require here are:

- feedback from students, lecturers, examiners, administrators, and expert judgement through questionnaire and interview on the washback effect of the test in the classroom

- stakeholder survey through questionnaires on the effect of the test on society, e.g. on end-users of test results

- statistical analysis (students' bio-data and psychological characteristics: to detect bias in the test for or against groups of students defined by these bio-data characteristics, such as gender, age, number of years studying or exposed to the language)

- expert judgement, student and lecturer opinions on differential validity

2.6 Criterion-related validity

This evidence refers to how the test relates to other external measures of the same ability. It reflects the consistency of several tests measuring the same ability or trait. To determine this we need

- comparison with future behaviour if possible, or with different measures of the same ability

2.7 Reliability evidence

This evidence deals with the consistency of measurement in the test and addresses the question "How much of an individual's test performance is due to measurement error, or to factors other than the language ability we want to measure?" (Bachman 1990). We need

- expert judgement of examiners and lecturers on the rating procedures (rater training, standardization, rating conditions, rating, moderation) and statistical analysis of scores awarded (mean, standard deviation, correlation, estimates of reliability, standard error of measurement, parallel forms reliability, inter-rater reliability, multi-faceted Rasch analysis to investigate intra-rater reliability, bias, use of the criteria, and similarity of the tasks)

- expert judgement of the examiners and administrators on the grading process (consistency and stability of the grading and awarding process; looking closely at

the performance of candidates who are close to the pass/fail boundary to ensure the fairness of the final results before they are issued to candidates).

Stage 3 Apply validation instruments and procedures to the existing speaking test

At this stage, piloted survey and other instruments and procedures from Stage 2 will be revised where necessary and applied to the existing speaking test instruments and procedures. The subjects will be selected from the same places as those in stage 2. More than ten thousand students will be taking this test twice a year as a normal part of their degree studies. A representative sample will be chosen from these.

Stage 4. Create a new computer-based speaking test in light of data from stage 3.

On the basis of evidence generated in stage 3, a new semi direct computer based version will be developed as follows:

- Establish weaknesses in the direct speaking test
- Find solutions to these weaknesses
- Develop specifications for a new semi-direct computer-based speaking test
- Develop a new instrument and procedures for a semi-direct computer-based speaking test through a core group of experts
- Trial on samples as in the piloting in stage 2
- Revise the test where necessary

Stage 5 Apply final versions of framework instruments and procedures to the new semi-direct speaking test. At this stage, we will

- administer the new test to a representative sample of students who will also take the existing direct test

- evaluate the new improved speaking test against the final versions of the framework instruments and procedures

All of the instruments described above are presented in the matrix (Figure 3.1) below according to the corresponding stages of the study. A research timescale follows the matrix; this is how the study is organized in terms of time, the stages involved and action to be taken by the researcher.

Figure 3.1 Summary Matrix: methods and research questions

Questions & Stages / Instruments	Questions 1 and 2				Question 3
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Literature review	X				
Expert judgement	X	X	X	X	X
Documentary data	X	X	X	X	X
Interview data	X	X	X	X	X
Questionnaire data	X	X	X	X	X
Observation		X	X	X	
Checklist		X	X	X	X
Statistical data		X	X	X	X

The next section provides a detailed account of how each research instrument above is developed, including the time each one takes to prepare.

SECTION II: Development of Research Instruments

The following tables (Figure 3.2 and 3.3) show the details of Main Study 1 (direct test validation) and Main Study 2 (semi-direct test administration & validation): each component in the framework is divided into several elements, and each element is evident (or not) as it is discovered by the instrument/source of validity evidence. Thus the report that follows shows the validity evidence (s) that prevailed from the direct speaking test, (and later in the chapter the semi-direct computer test), according to the instrument (s) used to gather these evidences.

Figure 3.2 Details of data gathering for Main Study 1

Stage of research	Instrument & its description	Participant	Number of *Respondents	Time of administration
MAIN STUDY 1 (Jan-Mar 2004)	**The speaking test**			
	Direct test with two parts & between 4 candidates:	4 Students		2 times a year in Feb/Mar; Sept/Oct
	Task A ~ individual Presentation	2 Examiners		
	Task B ~ group Discussion			
	Questionnaire			**Pilots Mar 03/ Sept 03; MS 1 Jan-Mar 04**
	Validation questionnaire:	Students	387	
	Student version: 3 parts	Lecturer/Examiner	46	
	(sections A,B,C)	Administrator	7	
	Staff version: 5 parts	Expert	9	
	(sections A - E)			
	Interview			
	Semi-structured interview	Students	64	
		Lecturer/Examiner	10	
		Administrator	6	
		Expert	5	
	Documents			
	Syllabus	Gathered by researcher	6 parallel sets of questions	
	Test specifications			
	Question paper			
	Test documents:			
	Score sheet, rating			
	criteria scoring guide,			
	instructions for			
	administering and			
	marking the test			
	Observation			
	Conducted on speaking test sessions as both observer & examiner	Conducted by researcher	as Observer 10x as Examiner 4x	

Figure 3.3 Details of data gathering for Main Study 2

Stage of research	Instrument & its description	Participant	Number of *Respondents	Time of administration
MAIN STUDY 2 (Sept 2005)	**CBT Monologue** Computer-delivered individual presentation	Students participate in multi-media computer labs	approx. 10	**Trials Sept 04 / Feb 05** **MS2 Aug-Sept 05** After student has taken direct test
	CBT Group			
	Computer- delivered discussion	Students participate in multi media comp labs	approx. 50	After student has taken direct test
	Questionnaire (for both direct test & CBT)			
	Validation questionnaire:		All students who had participated in CBT monologue and CBT group	Week of speaking test onwards
	Student version: 3 parts (sections A,B,C)	Students		
	Staff version: 5 parts (sections A - E)	Lecturer/Examiner Administrator Expert		
	Interview (on both direct test & CBT)	Students	all students who had participated in CBT monologue and CBT group	Week of speaking test onwards
		Lecturer/Examiner		
		Administrator		
		Expert		
	Document (for direct test)			
	Syllabus Test specifications Question paper Test documents:	Gathered from the university by researcher	6 parallel sets of questions	Week of speaking test
	Score sheet, rating criteria, scoring guide, instructions for administering and marking the test			
	Document (for CBT)			
	Specifications	Prepared by researcher		In progress prior to data collection dates
	Test script			
	Rating criteria			
	Observation	To be conducted on both direct test & CBT		Week of speaking test onwards

* Number of respondents will change updated as the study progressed.

Each research instrument is explained below in terms of its purpose, development and administration.

3.2 INSTRUMENTS FOR MAIN STUDY 1

A. The Speaking Test

The speaking test is a direct test, conducted face to face between four candidates and at least two examiners. Each term the exam is conducted for a week, involves most members of staff at the Language Centre in the main campus as well as in branch campuses across the country, students also in the main campus and the branch campuses, and six parallel sets of question papers. Students have a question paper and the examiner as stimuli; they read the instructions and situations for the exam in the question paper and examiners give them brief instructions before the exam begins. Each exam session takes up to 20 minutes to conduct; 2 minutes preparation time for each task (A & B), 2 minutes presentation time for task A, 10 minutes presentation for task B.

1. The Speaking test Questionnaire

Cohen et al (2000) suggest that one of the most important stages in the process of operationalizing a questionnaire is to take a general purpose or set of purposes and turn these into concrete, researchable fields about which actual data can be gathered.

The questionnaire for this study aims to validate the existing speaking test conducted in UiTM Malaysia, and more specifically, to gather data according to a validity framework which takes into account three major components (context, theory-based, scoring

validities) which defines the construct the test measures. It is intended to survey a large size sample of participants from the university, and thus, the design is a highly structured format, covering the elements found in the three validity components of the framework.

■ **Purpose**: The general purpose of the questionnaire is to collect data/feedback in an attempt to validate the existing speaking test that is conducted in UiTM Malaysia. Specifically, data/feedback is gathered according to three components of validity, listed below in the design section. All this information will be incorporated into the decisions made when the new computerized speaking test is developed in the second phase of the study.

■ **Respondents:** Students, lecturers/examiners, administrators (dean, deputy dean, course tutor, course coordinator) and expert judges, all from the main campus and the branch campuses of the university will be given the questionnaire.

■ **Design**: It is based on the socio-cognitive framework for validating a speaking test. It covers the three different validity types and the corresponding elements in each type:

- Context validity
- Theory-based validity
- Scoring validity

The questionnaire has 5 sections; each section contains statements that are ranked according to a 5-point Likert scale (1 for 'Strongly disagree' through 5 for 'Strongly agree'). There are two final versions:

- Student questionnaire : Sections A (Context validity), B (Theory-based validity), C (Scoring validity)
- Staff questionnaire: Sections A - C as in Student questionnaire, plus Section D (Consequential validity) and E (Criterion-related validity)

■ **Procedure:**

The first group of items for the questionnaire were first produced at CRTEC, Roehampton University during the 3 months period of November 2002 through Jan 2004; the main purpose was to translate or operationalize elements of validity in the framework into questionnaire items, which could elicit information from participants regarding these aspects of the speaking test. After several revisions, 78 items were decided upon, i.e. sections A (context), B (theory-based), C (impact on teaching & learning). However, before the questionnaire could be used in the main study, it had to be checked for language difficulty and any ambiguity in meaning. This was done at a very early stage before the pilot studies were carried out; in February 2003 the questionnaire was sent out for needs analyses by several students in Malaysia. Feedback obtained is listed below, and after changes had been made the questionnaire was ready for its first pilot study.

The following is a description in terms of the time and events that took place in the development and administration of the questionnaire:

1. Writing items for the questionnaire based on elements of the validity components of the framework. Items were written for each element in the framework; for example, two items were developed for 'purpose' and two for 'time constraint', three items on

language range, and so on, which were all elements of test 'task setting' in context validity.

Note: Items were generated based on theories of validation (see Ch. 2 for details) and according to the framework proposed by Weir (2005) and were produced by the researcher.

2. Discussions were held with 'expert panel' at the Centre for Research in Testing, Evaluation and Curriculum in ELT (CRTEC) at Roehampton University, to review and revise the items on the questionnaire.

3. Between Jan 8 - 20: Questionnaire was sent to Malaysia for a needs analysis on the linguistics aspects (vocabulary, structure, function)

4. Based on feedback from Malaysia, further analysis of the potential language problems in the student questionnaire undertaken, and changes were made accordingly.

Feedback on the questionnaire was received and the following points were noted:

a) Problem words/phrases were changed to simpler ones; some sentences were omitted, and others were retained (see Appendix 3.1 for details)

b) Items highlighted /changed:

Sec **A** 10, 18-20, 28-30, 37,41c, 42b, 43 a-c: all retained; response consistent

Sec **B** 12 –interesting strategies to note, 14-16: retained; response consistent

Removed: Sec **A**42a, Sec **C** 8

a) Pilot Study 1 (March 2003 in Malaysia)

After checking the questionnaire for language difficulty and ambiguity, and changes had been made on the items (e.g. open ended items were removed), pilot studies were carried out in Malaysia in 2003. The first of two pilot studies was conducted especially with the cooperation and support of members of the Language Centre at Mara University of Technology, Shah Alam campus.

Questionnaire used SET 1: 82 items

Total number of response: 80

The sequence of events for data collection is listed below:

STAGE 1: Administering the Questionnaire

1. Meeting with Deputy head (English dept), Academic & Administrative Coordinators of the Language centre on work concerning video recording of speaking test, and administration of the questionnaire survey to students. Letters were prepared regarding the above matter and sent to the parties concerned (Language Centre, Training & Staff Development Unit, the Registrar's office)
2. Recordings of the speaking test were conducted on the following groups of students from the faculty of Accountancy:

   AC 110/FT 3C - 4 groups

   ACA 110 (fast track) - 3 groups

Total time taken for recordings: Approx. 20minutes per group, total = 140minutes

3. Administering student questionnaire to the following groups of students,

after they had completed the direct speaking test:

Group: AC110/ Ft 3C; **n= 18**

Group: CS111/ B ; **n= 31**

Group: AC110/ B; **n= 25**

Group: CS113/ D; **n= 7**

STAGE 2: Data Analysis of March 2003 Pilot Study

The main reasons for performing the analyses were as follows:

a) To ascertain if the results obtained from the survey show an overall pattern, i.e. if participants had responded according to the intended purpose of the questionnaire

b) To establish the internal consistency of the items

c) To review items on the questionnaire in preparation for its use in fieldwork (Jan - Mar 2004), i.e. to sort, revise, & discard items where necessary

d) To obtain results that can provide useful information regarding the existing speaking test conducted at UiTM (weaknesses, strength, etc.); analyses to be done on SPSS: Descriptive statistics, Reliability analysis and Factor analysis

♦ FINDINGS/DISCUSSIONS

SPSS data analysis (see Appendix 3B in attached CD I)

Descriptive statistics results:

- Response for majority of the items is above 3 (undecided); only five items had a mean rating of below 3, i.e. items 12, 29, 43a, 43b, 43c

- Mean standard errors for all items are low; standard deviations are low as well
- Frequency table shows how students responded to each item; the five items rated below 3 (i.e. 2=disagree, 1=strongly disagree) need looking at, especially item 43
- Descriptive Analysis data points to the fact that while most respondents chose options 4 (agree) to 5 (strongly agree), there are also those who selected options at the other end of the scale: for items a7, a11, a12, a18, a20, a26, a28, a29,a33, a36, a37, a40, the cumulative frequencies for options 3, 2, 1 are higher than 60%.

Items a7 – a20 are related to item specifications, while a26 – a40 are regarding examiners. It appears that there are several aspects of the test that students are not adequately informed about, or not informed about at all, such as criteria for marking, test types, test functions, exam setting, etc.

Internal Consistency analysis results:

- Cronbach alpha reliability coefficients is used as it is based on the average correlation of items within a test if the items are standardized; if items are not, it is based on the average covariance among the items
- Our analysis shows alpha = .7201 and standardized alpha = .7889
- There were 4 items with negative item-total correlations; there were 7 other items with very low item-total correlations → If these items were removed, the alpha coefficients would increase (see Reliability analysis below)
- Internal Consistency Analysis of our data shows an alpha that is quite high;

However, items that had very low and negative total correlations might indicate that students were, again, not familiar with these facets of the test content, and hence were not confident of how to respond to related items in the questionnaire.

Factor analysis results:

- After some reorganization of items in the questionnaire, the FA output shows items that are loading on the same factors; some patterns are emerging .

  A closer look at each group of items provides possible explanations.

  For e.g. items a3, a4, a5, a6, a7, a39 are loading on factor 1; except for a7 which has the lowest index, the others are related to what students do/experience in the test

- Factor Analysis data showed that some items load together on one factor, while some others do not. In relation to student responses, again, 17 factors could have influenced this. Items that are rather disparate in this analysis might point to several possible reasons such as negative washback effects of teaching, time constraint, students are not adequately informed about test, and so on.

♦ Comments on the existing speaking test

All analyses were based on the students' response to the questionnaire. It is crucial to note at this point that since this is the first time such a survey had been conducted on the test; factors that cannot be explained by our data could certainly affect the results (see Cohen et al 2000: 253-255)

- In general, students' responses were positive yet inconclusive, there is some pattern emerging in our reliability and factor analysis data, and it appears that there are several aspects of test content that students are not clear about
- Based on the findings above, the following decisions were made:

▪ After a careful study of the descriptive statistics data (Frequency output), items that had a high percentage (>60%) of 'undecided' - 'disagree' options were removed in stages from the questionnaire

Note: This was done as they may not function well in the set of items and as a result distort the data.

▪ Reliability analysis was done each time items were removed. After several runs, alpha increased as the number of items decreased, as is shown in Figure 3.4a:

Figure 3.4a Reliability analysis of questionnaire items

Alpha	Items
.7620	51
.8228	36
.8320	31
.8406	29

▪ Factor analysis was conducted and the final output on 37 items (Section A) showed the results in Figure 3.4b below. Five items appeared insignificant or had negative to zero index : items 7, 15, 21, 31, 17d

Figure 3.4b Factor analysis of questionnaire items

Component that items were loading on	Item	Possible factor
1	41a, 41b, 41d	test setting
2	3, 4, 5, 6, 39	familiarity with task
3	27, 32, 34, 35, 42c	Interaction with examiner & other speakers
4	1, 2, 22, 24, 25	language of task
5	30, 38, 43c, 17b	wording of these items
6	19, 20, 23	information in text
7	11, 14, 16, 25, 42b	Task
8	42a, 17a, 17c	test preparation

♦ Conclusions to above findings:

Although the data for Reliability and Factor analyses were computed using fewer items from the original questionnaire, the following decisions were made:

1. **First revision** of the questionnaire:

   Items *removed*: open-ended items (Sec A: 10, 41e / Sec B: 12, 26)

   Second revision of the questionnaire (Sec A only)

   Items *removed*: 8 & 9 (collapsed to item 11); 43a, 43b (data showed respondents disagreed least with 43c)

   Items *revised*: 12, 13, 18, 19, 20, 22, 26, 28, 29, 33, 36, 37, 40, 41c, 42a, 43.

2. The data gathered from our analyses helped as an indicator of the internal reliability of the items and how the respondents performed in the survey, but it

doesn't provide us with a complete picture of the survey. However, the following points were clear:

- Data helped us see some patterns emerging
- Revisions could be made progressively on the questionnaire as more data is collected and analysed
- Data analyses, item changes, and item revisions would be recorded periodically

STAGE 3: Final Discussion on Findings

Based on the analyses and findings above, discussions were held and the following points were decided upon:

1. Minor changes were to be made to *Sec A*: numbers 8, 37
2. *Sec A* would then be ready for piloting on UiTM students
3. *Sec B & C* required several adjustments; all sections are now ready to be sent out for Pilot study 2: Student questionnaire (sections A, B, C)

b) Pilot Study 2 (September 2003 in Malaysia)

As in Pilot study 1, this second pilot study was conducted especially with the cooperation and support of members of the Language Centre at Mara University of Technology, Shah Alam campus. We also managed to extend the survey to participants from a branch campus (Jengka campus in the state of Pahang)

Questionnaire used SET 2: 100 items

Total number of response: 65

STAGE 1: Administering the Questionnaire

Student questionnaires were administered to participants as follows:

Campus	*Course*	*Number of respondents*
Jengka	AS117	**12**
Shah Alam	AP116 A	**24**
Shah Alam	AP116 B	**8**
Jengka	AS117	**21**

STAGE 2: Data Analysis of September 2003 Pilot

The results of data analysis (questionnaires) were as follows.

SPSS data analysis (see Appendix 3C in attached CD I)

a. Descriptive statistics The following results were apparent:

Sec A: The following items had means < 3.00 (closer to disagree/ strongly disagree)

Item 7 (Both tasks A & B should have equal marks) ~ **75%** disagreed

24 (Effect of examiner accent) ~ **62%** disagreed

→ Examiner accent does not affect their performance

25 (Prefer same gender examiner) ~ **70%** disagreed

→ Gender is not an important factor for performance

40 (Computer-based speaking test) ~ **75%** disagreed

→ Computerized speaking test is not favoured

Other items in this section where results were skewed:

Item 27 (Acquaintanceship with examiner) ~ **93%** agreed

→ Knowing the examiner is important

29 & 36 (2nd examiner's presence) ~ **55%** disagreed

→ The presence of 2nd examiner is not favoured

Sec B:

Items 3 & 19 (How to satisfy examiners) ~ more than **55%** disagreed

→ Not considered at preparation stage

8 & 18 (Knew enough specific info from previous....) ~ **51%** disagreed → Insufficient internal/background knowledge on subject matter

14 (Ideas presented smoothly) ~ **45%** disagreed

15 (Ideas presented in organized fashion) ~ **65%** disagreed

→ In ability to /lack of confidence in presentation for task

31 & 32 (Ideas presented smoothly and in organized fashion) ~ **60%** agreed

→ Able to/have more confidence in group discussions for task

B; better marks were obtained here than in task A

♦ Conclusion to above findings:

Apparent factors that influence performance in the speaking test:

a) The examiner:

- Gender
- Acquaintanceship
- Presence of 2nd examiner

b) Processing the task:

- Satisfying examiner is not a consideration
- Students not successful in accessing internal/background knowledge for both tasks
- Students more confident in presenting the task smoothly & in an organized fashion for task B than for task A

b. Internal Consistency analysis

- Cronbach alpha reliability coefficients are used as they are based on the average correlation of items within a test if the items are standardized; if items are not, it is based on the average covariance among the items

Note: the minimum acceptable alpha is .70

The analysis shows the following:

SEC A: alpha = .8094

- There are no items with negative item-total correlations

- There are five items that have alphas that are not significantly lower; only one that had alpha .7978; however, since the difference between this and the average alpha is only .01, the item is retained

SEC B: alpha = .8090

- No items with negative correlations

- seven items with alphas below .80, but all of them close to .80: lowest .7945 highest .7999; however, since both have differences of less than .01 from the average alpha, they were retained

SEC C: alpha = .8407

- No items with negative correlations
- Only two items had alphas lower than .83: C6 - .8178, C7c - .8179; both items were retained because the difference between them and the average alpha is only .02

♦ **Conclusion to above findings:**

This analysis on the Sept pilot data shows alphas that are acceptable for all sections. Like data from March 2003 pilot where there were fewer items, candidates' responses seemed to result in a high correlation of items in the questionnaire. However, unlike the last pilot study where there were negative & very low correlations amongst items, this data showed positive results in terms of the quantitative output.

c. Factor analysis

- It was noted that items in this questionnaire are not typical of attitudinal type questionnaires where factor analysis would produce results appropriate for analysing the underlying construct of these questionnaires. Items in this questionnaire appear to be concrete/factual in nature, especially in the section on context validity; the results can only be used as possible indications of the way candidates responded to the items, and possibly show emerging patterns in the response

- The analysis showed that some items loaded together on one factor, while some others did not:

SEC A:

- After several runs, the data showed items loading on a maximum of 16 factors (80.3%variance explained)
- While this model resulted in a major proportion of the variance being explained, it did not seem to offer an easily explained set of factors.
- As a result further analyses were performed, first with an 11 factor model (which explained 68%variance explained), and finally with an 8 factor solution (which explained 58% variance explained)

Figure 3.5a Factor analysis on questionnaire items Section A

Items	Factor
1, 2, 3, 6, 11, 19, 20, 23, 37b, 38b	External conditions/instructions
8, 25, 26, 28, 32, 33, 34, 35, 39a, 39b, 40	Knowing other speakers' conditions
29, 30, 31, 33, 36	Interlocutor conditions
4, 5, 12	Communicative intention
16, 17, 38c	Type of information
item 7	Weighting
item 37d	Physical condition
item 28	Acquaintanceship

SEC B:

- After several runs, the data showed items loading on a maximum of 16 factors (81%variance)
- While this model resulted in a major proportion of the variance being explained, it did not seem to offer an easily explained set of factors.

- As a result a further analysis was performed with a 10-factor model solution (which explained 59%variance).

Figure 3.5b Factor analysis on questionnaire items Section B

Items	Factor
11, 14, 15, 25, 26	Presentation (language/style)
16, 17, 18, 19, 33, 34	Preparation (with others)
2, 9, 23, 32	Interpret information
5, 12, 35, 36, 37	Speech functions
6, 8, 22, 28	Use of knowledge
27, 38, 39	Conducting discussion
1, 3, 30, 31	Learned strategies
7, 13, 21	Topic
4, 20, 27c	Mental notes
10, 24	Self-adjustment

SEC C:

- After several runs, the data showed items loading on a maximum of 13 factors (100%var)
- While this model resulted in a major proportion of the variance being explained, it did not seem to offer an easily explained set of factors.
- As a result a further analysis was performed with a 4-factor model solution (which explained 80.7% variance)

Figure 3.5c Factor analysis on questionnaire items Section C

Items	Factor
4, 5, 6, 7a - 7e, 9	Knowledge used/practised
2, 3	Discussion in class
1	Information given in class
8	Knowledge used/practised

♦ **Conclusion to above findings:**

Though it was stated earlier that items in this questionnaire are less suitable for use with factor analysis, the results indicated **emerging patterns, especially for sections B & C.** All items in both sections loaded on some factors; however, this was not the case for section A. It was also noted that the items grouping together on a factor, in fact, reflected upon the elements of the framework for speaking, for e.g. items B16 - 19 loading on the factor called 'preparation with others' are reflective of the elements of conceptualizer/goal-setting in the first stage of internal processing for a speaker.

CONCLUSION

- Analyses on our data helped us see some patterns emerging: this is apparent in frequency statistics, reliability analyses, and factor analyses
- Further revisions will be made on the questionnaire and this will be trialled once more before Main Study 1 commences in Jan 2004
- Some questions raised in this analysis were incorporated into the design of the Interview instrument (see notes on *The Interview*). These are questions/items relating to scoring validity of the test, discourse mode & nature of information in the test task, examiner's role, and computerized testing.

♦ **Final report on the Questionnaire for Speaking**

The questionnaire was designed during the period of Nov 2002 - Mar 2003.

(see Bell 1987; Brown 1988; Foddy 1993; Cohen et al 2000; Robson, C 1993/2002; Booth et al 2003)

- It was designed for use as an instrument for data collection in Main Study1
- It was based on a framework for validating a speaking test which was grounded with theories on test validation, speaking as a construct, and cognitive-processing of the construct (Weir 2005), contains five sections, based on validity components of the framework:

  Section A: Test content (Context validity)

  B: Internal processing (Theory-based validity)

  C: Scoring of the test (Scoring validity)

  D: Effects of the test on teaching & learning (Consequential validity)

  E: Criterion-related validity

- It was piloted on students at Mara University of Technology, Malaysia, on two occasions: Mar 2003 and Sept 2003
- It was presented at the Language Testing Forum 21-23 Nov 2003 in Cambridge; expert opinions/comments/feedback were obtained on reliability issues of items, the design of the study for test validation, and issues related to computer delivery of the test.
- The versions prepared to date are:

1. Student Questionnaire (sections A, B, C): for student participants
2. Staff Questionnaire (all sections): for other participants, i.e. lecturers, examiners, administrators, experts
3. The Speaking questionnaire (all sections)

(see Appendix 3.2 + 3.3 for student and staff questionnaires respectively)

2. The Speaking test Interview

One of the unique features of the interview technique is that it involves knowledge generating between humans, often through conversation, which is a marked difference from other research methods which sees human subjects as manipulable and data as external to individuals (Cohen et al 2000). The purposes of the interview are also many and varied but for this study, the main purpose is to gather firsthand information from various participants who are directly and indirectly involved in the speaking test, about the test. For this purpose, the interview used in the current study is less formal, a guided yet open-ended approach where topics, questions and their sequence are determined in advance, but the interviewer is free to modify the sequence of questions, change the wording, explain them or add to them as she sees fit in the course of the interview.

- **Purpose**: The purpose of the interview is to collect data/feedback on the existing speaking test that is conducted in UiTM Malaysia. It attempts to gather participants' views and thoughts on the existing speaking test, based on:

  a) the framework for validating a speaking test

  b) participants' response to the questionnaire surveys conducted in the

earlier pilot studies

As with data from the questionnaire, data/feedback from the interviews will be incorporated into the decisions made when the new computerized speaking test is developed in the second phase of the study. This is the first set of *qualitative* data collected for the study.

- **Respondents:** Students, lecturers, examiners, administrators (dean, deputy dean, course tutor, course coordinator) and experts (lecturers & other faculty members who have wide experience and knowledge in the field), all from the main campus and the branch campuses will be interviewed at various times and locations, especially after the speaking test had been conducted.

Participants will take part in the interviews as follows (see Sec B. Research Proposal, pp 17-22):

1. Students

Stage 1: 1.2, 1.3

2: 2.3, 2.5

3: on direct speaking test

5: on new computer-delivered test

2. Lecturers/Examiners

Stage 1: trial interview on direct speaking test

Stage 2: 2.5, 2.7

3. Administrators

Stage 2: 2.5, 2.7

4. Experts in the field

Stage 2: 2.5, 2.7

■ **Design**: The trial interview takes on a semi-formal, semi-structured format, leading to a more structured interview where the items/questions are generated in line with the framework for validating the speaking test and the actual test format. It is divided into 3 stages of: questions on test preparation, questions on presentation, and general questions such as on test fairness and test difficulty.

■ **Procedure:** Before the interviews could be conducted in the main study, it was trialled to check for language difficulty, ambiguity in meaning, and types of response. All responses were recorded and transcribed for analysis. Though the trials enabled the researcher to make some conclusions about the interview technique, the participants, and their response, further work was needed to enhance the quality of the instrument. The process was conducted as follows:

1. Interview trials conducted at UiTM Jengka & Shah Alam campuses
2. Interview data from September trials were analysed and compared with questionnaire items.
3. Discussions were conducted on the data, objectives of the study, and final format of the interviews. As a result, the structure of the interview was formulated and questions for the interview were revised and re-written.
4. Based on findings from the September trials, it was decided that the following items were to be included in the interview:

   a) scoring validity of the test

   b) discourse mode

   c) nature of information in the test tasks

   d) examiner's role

e) computer delivery of test

a) Interview trials (transcription): September 2003

All interviews were carried out through the medium of English, a decision that was based on the fact that all interviewees were bilingual speakers who were comfortable with both English and Malay. The transcriptions are included below as a sample of what transpired during the sessions. As will be noted later, the transcriptions for interviews from the main study will be included in the appendix section of the dissertation (see Appendix 3.4 + 3.5 for student and staff interviews respectively).

12Sept 2003
Participants: 4 students Diploma Science
Campus: Jengka, Pahang
Interviewer (Intw)
Participant A
Participant B
Participant C
Participant D

Task A
On preparation:
Intw: Do you remember what you did to prepare for the task?
A : I thought of the points, jot them down in complete sentences...but when I presented, I didn't use them....I thought of other points instead
B : I did the same thing but had trouble presenting the points..
C : I studied the topic, understood it & jotted down my points
D : I'd think in Malay...translate into English, jot the points in English...also in paragraph form

On presentation:
Intw: What about the oral presentation?
A : I prepared my speech by writing it down, but had major vocabulary problems during the presentation...can't think of the right English words to use
C : I had major problems with the language use...tried some self-corrections, but I think I had an advantage being the last person to present...I listened to the others & used some of their ideas/words...
D : Major problems with both language and vocabulary...also added on other points during the pres... I didn't have enough time to say everything...

Task B:
C : We prepared individually then only presented...I think the instructions for the task were vague/confusing... I was less nervous here but nervous overall, especially because others were in the room...other students, and 2nd examiner...
(Other participants mentioned the same points)

Intw: Did you all feel that you had enough practice before the test?

C : Yes, but the format was slightly different from the one given in the test; during practices, points were given, in the test you think up your own points to present

D : I also felt nervous especially because other students were around
A : Same with me.., but the group presentation was ok...I contributed the least though...
B : I said even less, was nervous & let the others speak since their English were better than mine

16Sept 2003
Participants: 4 students Diploma in Architecture
Campus: Shah Alam, Selangor
Interviewer (Intw)
Participant A
Participant B
Participant C

Task A
On preparation:
Intw: Do you remember what you did to prepare for the task?
A : Yes, the question was clear...so I jotted my points down, they are based on my experience...I included a greeting for introduction and a conclusion, as we learnt in class...
B : I wrote points down accordingly; I planned the presentation such that I included some general knowledge on the topic

Intw: So, the topic was familiar to you?
A & B : Yes, we know a little about it... about social problems among youths
C : I had a different topic, I think it's a bit harder... but the instructions were quite clear
A & B : 2min is sufficient time to prepare
C : I feel it's not enough, but I was lucky there were visitors who took up a bit of the time!

On presentation:
A : I feel that the candidate you are, A, B, C or D, makes a difference because the later ones have advantage of more time to think... Anyway, I was nervous, & lack of points... could've done better, but language was not a major problem, & I made some adjustments/corrections in my speech
B : I spoke too fast so had extra time... had to think of other points to say; my language was quite poor due to nervousness...and, I think my task as candidate D was quite difficult compared to others'
C : I spoke too fast too, and had extra time but couldn't say any more... I think overall I was ok

Task B
On preparation:
Intw: Do you remember what you did to prepare for this task?
All three: You are not allowed to discuss during the preparation; prepare on your own,

discussion is during the presentation
C: Group discussion seems easier though the others (other 2 from another class) didn't help...10 minutes is too long

A & B : We were in the same group, and yup, the other 2 from another class didn't help toward the discussion; all they said were they "agree" or "disagree"

Intw: Should we reduce the number of people in a group then?
All three: Not really... The examiners helped those who were too quiet & didn't say very much... Language problems were clearer in Task A than B

17Sept 2003
Participants: 4 students Diploma in Architecture
Campus: Shah Alam, Selangor
Interviewer (Intw)
Participant A
Participant B
Participant C
Participant D

Task A
On preparation:
Intw: Do you remember what you did to prepare for task A?
A : I jotted my points down, including words that I might use, & words that are different...I added notes as well
B : I considered the task, wrote down a number of points, and the introduction
C : I thought about the points carefully, i.e. why I would say something, also prepared introduction, body, conclusion
D : I jotted down points too, but had difficulty in making them coherent...

Intw: Did you have enough time to prepare?
A : Not enough, especially because topics on current issues are difficult & big!
C : Since I've taken the MUET, I find that the time given is manageable...
B : I think it's enough for me...
D : Not enough time to prepare...

On presentation:
D : My nervousness and anxiety affect my language use
C : Not a serious problem, just certain word usage...vocabulary
A : Because there is a lot to say, I was worried about accuracy... I became spontaneous in my speech & stated the points not according to how I recorded them... just said what is best at the time...
B : I had some grammar problems...

Intw: What about the presence of the 2nd examiner?
All four: Not a problem because we knew her already... in fact, we knew everyone who was in the room... quite important that we know them...

Intw: So do you think this was a fair test of your speaking ability?

C : Overall, it's a fair test, but can be better, especially the topics and time for preparation; it's a real problem economizing time, what to do 1st, 2nd, and so on...But in the test today, I even thought of idioms to use!
A : Problem is need to prepare better notes, and to focus on a few points but elaborate them well...it's quite a fair test
B : Just glad to finish the test, as long as there's enough practice, and some experience & imagination!
D : Don't feel good; in task A I didn't use all my points, in B, I spoke very little...

♦ **Points raised in the interviews**:

The following were points that emerged from the interviews, and it was interesting to note how the groups differed in their ideas and strategies for the test.

- Weaker students

Preparation stage:

1. List/jot down points; some in complete sentence structures
2. Have clear problems in:
   - Content knowledge
   - Topic familiarity
   - Language use: use of L1 for thinking through & writing points during preparation

Presentation stage:

1. Major problems with language use for
   - expressing ideas in complete sentences/ structures
   - completing ideas
2. When stuck with an idea/expression, tendency to just stop or quit speaking; very few attempts at corrections or adjustments
3. Highly dependent on notes, and other speakers, i.e. listen to others' presentation to get help for ideas or language use

4. Speak less than what's been prepared; difficulty in being spontaneous, & voluntarily let others speak

- Proficient students

Preparation stage:

1. List down points
2. Incorporate background knowledge & experience into notes
3. List specific structures & vocabulary to use, and occasionally a plan for the speech (introduction, content, conclusion)

Presentation:

1. Minor/no problems with language use for completion of & coherence in ideas
2. Able to speak spontaneously; included ideas that were not listed in notes during preparation
3. Less dependent on notes but more concerned with accuracy of speech
4. Able to make corrections or adjustments during presentation
5. Very concerned with outcome of presentation, i.e. thinking of points to focus on & elaborate, rather than listing of many points

♦ **Overall comments from candidates:**

1. The candidate you are assigned affects performance, i.e. whether you are candidate A, B, C or D. The fact that the topics & points assigned to each candidate had been pre-determined affects a student's performance.
2. Topic: in terms of difficulty, quite different between groups, some topics more difficult than others (e.g. sets 5, 6 are difficult). There seems to be an agreement

here that topics are not equivalent for all groups; hence, those with more challenging topics are at a disadvantage especially in terms of time for preparation.

3. Preparation time of two minutes for both tasks is insufficient. This is also related to point 2 above. Looking at the topics & demands of the tasks, two minute preparation time does seem short for a candidate to be able to prepare a well thought & organized presentation sufficiently.
4. On test fairness, most candidates were undecided because:
    a) preparation/practice in class took place very late, about two weeks before the test; thus they were not prepared, especially in terms of anxiety leading to the test date
    b) topics during the test seems more difficult than the ones used for practices; in some cases, even test format seemed different, unequal between classes
    c) There seems to be differences between classes & lecturers in terms of preparations leading to the test

b) Based on the trials above, the following decisions were made:

- To gather feedback on the speaking test, after analyses of the questionnaire data, especially for items where response were not clear
- To conduct interviews with students especially for information on strategies they employ or processes involved in performing the tasks

- Tabulate comparison between interview questions and corresponding items in the speaking questionnaire
- Interview format for experts, administrators, lecturers, and examiners is semi-structured; interview format for students is open-ended

Figure 3.6 below shows the comparisons between interview items and their equivalence in the questionnaire. This is necessary in order to determine the items that need to be included in interviews (not everything already in the questionnaire), based on feedback from pilot studies; the structure would be an open-ended/semi-structured format to a more structured format

Figure 3.6 Comparison between interview items and questionnaire items

Interview items	Questionnaire items
1. What did you do to prepare for task A?	B. 1-9
2. Did you have problems in terms of: - Time for preparation? - Topic? - Language used in task? - Instructions for the task?	 A. 18, B. 7 A. 15-17, 19, 20 A. 1, 2, 13
3. What did you do during the presentation?	B. 10-15
4. Did you have problems in terms of: - Time for presentation? - Organization of ideas? - Coherence of ideas? - Presence of other speakers?	 A. 10, 11 B. 15 B. 14 A. 28
5. What did you do to prepare for task B?	B. 16-23
6. Did you have problems in terms of: - Time for preparation? - Topic? - Language used in task? - Instructions for the task? - Method of preparation?	 A. 18, B.21 A. 1, 2, 15-17 A. 13, 19, 20 B. 16-18
7. What did you do during the discussion?	B. 24-32
8. Did you have problems in terms of: - Time for presentation? - Interacting with other speakers?	 A. 11
Language required for group discussion? (stating ideas, justifying, agreeing, disagreeing, etc)	B. 27, 28, 29, 31, 32
9. Were you affected by the presence of a second examiner?	B. 24, 25, 26, 30
10. Did you get enough practice on individual presentations and group discussions in class before the test?	A. 29, 36
11. Overall, was this a fair test?	C. 3, 4, 5

It appears that most, but not all items in the questionnaire had been included in the interview. Based on this point, and the findings and data analysis of the Sept 2003 pilot study, it was decided that the following topics were to be included in the next revision of the interview items:

a. scoring validity of the test
b. discourse mode
c. nature of information in the test tasks (context validity: task demands)
d. examiner's role (context validity: interlocutor)
e. computerized testing (context validity: administration)

and all others according to the framework for validating the test.

CONCLUSION

Based on all the above information on developing the interview technique, data gathered from the September trials, analyses, discussions, revisions, and a final comparison of the outcomes of both instruments (interview and questionnaire), the 'Guidelines for the Interview' were developed (see Appendix 3.6 for a detailed account of the guidelines).

In summary, both the questionnaires and interview framework were developed as a result of theories on validation and research methods, and the pilot/trial studies conducted by the researcher at different stages of the instrument development. They were initially based on the socio-cognitive framework for test validation, piloted several

times, and the final forms emerged before Main Study 1 was conducted in Jan-Mar 2004 (details of the study & its findings are found in Chapter 4: Direct test validation)

3. The Test Documents:

A. Syllabus

B. Test specifications

C. Question papers

D. Score sheets

E. Guidelines/Instructions for administering and marking

F. Rating criteria/scale

The documents above were gathered by the researcher from the university in Malaysia where Main Study 1 and Main Study 2 were conducted. They can be found in the Appendix 3.7a – f, but each one is described below.

A. Course syllabus (see Appendix 3.7a)

Students who participate in the study are enrolled in an English course as follows:

Mainstream English II

Code: BEL 250

Level: Intermediate/Advanced

Components: Reading, Writing, Listening, Speaking

The objective of the Speaking component is to focus on training the students to plan, organise and participate in effective individual presentations and group discussions. ME II also prepares students to meet the requirements of the Malaysian University English Test/MUET; (for the objectives of the Speaking component in MUET see

booklet MUET Regulations and scheme of test, Syllabus and Sample questions; Malaysian Examinations Council 2001)

Note: The descriptions of the speaking component in the university syllabus reflect the objectives listed in the speaking component of the MUET booklet.

B. Test specifications (see Appendix 3.7b)

Like the specifications for other components, the one for speaking is divided into:

- Test objectives
- Text type
- Propositional features
- Organisational features
- Language level
- Weightage of marks & duration
- Marking guidelines

For example, the propositional features include familiar, academic and/or social, general topics related to family, college and community issues; the language level is formal/semi-formal; the duration is 2minutes preparation time for both task A & task B, and 2 minutes presentation time task A, 10 minutes presentation time for task B.

C. Question papers

There are six sets of the Speaking paper each term and examiners are to follow the order of when each set is to be used strictly. For example, students doing the exam on Monday will do only situations 1 & 2, Tuesday 3 & 4, and so on. Topics range from

general knowledge of current issues to events or activities that students are familiar with or have experience in at the university; for example, issues related to the December 2004 tsunami, reality television programmes, programmes for teenagers and the quality of graduates.

Note: This system is to ensure test security because the test is conducted over the period of a week, not only across various faculties in the main campus but also across branch campuses throughout the country. Confidentiality and strict conduct of the exam administration is vital so students and/or other members of staff are not able to obtain the topics before and during the exam. (see Appendix 3.7c for sample papers)

D. Score sheets (see Appendix 3.7d)

E. Guidelines/Instructions for administering and marking (see Appendix 3.7e)

F. Rating criteria/scale (see Appendix 3.7f)

* These test related documents are distributed to examiners as a package

* The score sheet for Speaking contains candidate identification, breakdown of marks for individual task (A) and group task (B), total marks and the final score after conversion to 15%. Each sheet allows for up to eight candidates or two groups, hence, each examiner may possess any number of the sheets, depending on the number of groups each one is responsible for.

* The instructions for administering the exam and marking are found in a document drawn up by the Language Centre. It contains a set of nine instructions and key notes on confidentiality of the exam and recording of the marks. To date, the instructions have not changed and these are distributed to all examiners every year; the only change would be the order of the six situations for the exam.

* The rating scale for Speaking contains three criteria and each one is marked according to a scale as described below. There are two sets of scale, one for the individual presentation and another for the group discussion task; however, the only difference is in the description of 'Communicative ability'.

Criteria for rating:

Task fulfilment : scale 1 (lowest) to 6 (highest)

Language ability: scale 1 (lowest) to 6 (highest)

Communicative ability: scale 0.5 (lowest) to 3.0 (highest)

Each criterion is described within the scales, for example:

Task fulfilment: a score of 4 > Fulfils task satisfactorily

Language: a score of 4 > Displays satisfactory control of language

Communicative ability: a score of 2 > Shows ability to contribute to the discussion Satisfactorily

Note: Discussions on the existing scale and the criteria used for rating the direct test were carried out between the researcher and examiners, lecturers and other experts at the university and CRTEC. Further details on issues related to the rating scale are found in chapters 4 and 5 on findings for Main Study 1 and Main Study 2 and for issues related to scoring validity.

4. Observation

Observations were conducted from the earlier stages of the study during the pilot studies and throughout the main studies by the researcher. During the observations, the researcher's main aim was to observe if the test was administered and marked according to the guidelines/instructions provided for context validity, and to some

extent how students behaved during the exam for theory-based validity. In total the researcher made approximately twenty observations within eight weeks of the Main study 1 phase of the research.

3.3 INSTRUMENTS FOR MAIN STUDY 2

Developing instruments for Main study 2 involved two stages: a pre-trial of the monologue computer test, and a pilot study of the first complete two-part computer test; these were the first times the computer test had been administered, both times at the university in Malaysia. In both cases, considerations had to be made for conditions for delivering the test (facilities/infrastructure, participants, scheduling, technical assistance etc.) and specifications for the test (format, topic, language functions, etc. parallel to specifications of direct test).

The design and development of research instruments for the direct and indirect tests had been described in detail in Section II above. The direct test, which was already in existence at UiTM Malaysia, is administered twice a year in March and September (also detailed in section II) by the Language Centre, and data for the present study was collected during the periods when the test was administered.

Hence, this section will detail procedures in which the indirect computer-delivered test, which is a prototype never used before, was established and administered.
We start with the pre-trials of the monologue computer test, in which several variables had to be established and are detailed below.

A. Pre-trials (Sept 2004) of the computer- delivered monologue test in Malaysia

The individual MONOLOGUE test (see descriptions below) was used for the trials in Malaysia. The main aim of the trials was to establish the possibility of administering such a test in terms of infrastructure, support and student ability. In establishing the minimum requirements and conditions for delivering the computer test, the researcher was involved in the following activities:

1. Focus group meetings with representatives from public & private institutions were held to discuss the test. These include the following institutions:
    - Mara University of Technology (public)
    - National University of Malaysia (public)
    - National Electrical Board Malaysia (private)
    - International Educational Consortium Malaysia (public)
    - Pointflex Malaysia Incorporated (private)

2. The meetings centred on the following subjects:
    - An account of the present PhD study on 'Testing of Speaking Using Computer Technology'
    - Issues surrounding the direct method for testing of speaking, with special reference to the MUET speaking test; this is important as all students pursuing a degree in the respective universities had/would have taken the test. These include concerns such as test practicality, context-based elements such as equivalent forms, interlocutor variables, and test reliability

- The monologue sample speaking test and whether it has the potential to address the problems posed by the direct test
- Technological requirements for such a test
- Some introductory information about the potential 'seminar' type task for the second task (group)

3. The trials for the computer test were conducted within the month in four centres and involved the following
   - approximately 100 participants, 4 technical assistants (at three computer centres/labs), 5 lecturers, the researcher
   - Each session involved a participant responding to a topic (4 sets of questions), after given from 0 - 60 seconds preparation time; speaking time is 1 -2 minutes each time. At the end of each topic, the participant responds to a questionnaire regarding what they thought of or did before starting the test, during preparation time and speaking time (theory-based validity)

4. The following were some **problems** faced during the recording sessions:
   - Availability of labs/scheduling problem
   - Recording errors especially in cases where there were too many candidates in the lab (25-30); some were not able to record all the tasks, but very few could not record any task
   - Technical problems related to computer set up, recording, and so on

Note 1: Overall, it was concluded that the outcome of the trials was positive and

encouraging. There is potential for the computer test to be conducted in Malaysia, as resources in terms of candidates, infrastructure and other facilities were accessible and support from various institutions was available when needed.

Note 2: All recordings were organized at CLARe, UK; they were listened to again, scripts of those poorly recorded & missing recordings were removed. Finally, of 100 recordings, 96 were used for editing & then sent for rating by experienced IELTS raters at the University of Reading

1. The Monologue Test

(see Appendix 3D in CD I attached)

The first test which was developed and delivered by the computer was a monologue test. It required the candidates to read the instructions in print, record their response into voice recorder software, and finally respond to a theory-based validity questionnaire which contain items related to what they did while preparing and speaking in the task. The questions in the test were adapted from the University of Cambridge IELTS speaking test. This monologue type presentation is similar to task A (an individual presentation) in the existing direct test.

The following is a summary of the test:

- An individual task which requires the speaker to speak on a given topic and for a specific time
- Test was developed using the IELTS task & format
- Four tasks were used, each with varying conditions, for planning time and

speaking time

- Each task is followed by a questionnaire on planning, thought processes, and Speaking (theory-based validity)
- Candidates' speeches were recorded on the computer using soft wares for voice recording such as DIVACE & Sound Forge 7.

2. The Semi-direct 'Interactive' task

(see Appended CD II for a copy of the test)

The next stage was to develop task B (discussion) as is found in the existing direct test. The proposed format was developed over a period of three months, and the final test was to be completed and used for Main study 2 in January 2005. The following is a liost of stages involved in developing the computer-delivered 'interactive' tasks.

a) Preliminary design and trials

August 2004:

Several clips of the direct speaking test were viewed (8 clips in total); the observation checklist for interactional speech functions (O'Sullivan, Saville & Weir 2002) was used to determine functions elicited by students during the group discussion task.

It was decided that a format that would be appropriate at this point was a 'semi' interactive task where the test taker will respond to a question posed during a discussion that had already began, i.e. test taker views/listens to the discussion on the computer screen & responds to questions/prompts presented at several points throughout the discussion. (see Appendix 3.8 *Sample script task B*)

September – October 2004 (in Malaysia):

Two groups of 3 students each participated in a video recording of the proposed group discussion test using the sample script; recordings were burnt into CDs

- To be edit in accordance with the sample script prepared (see *Sample script task B)*
- To be tried out with students (non-native speakers of English)

b) The Speaking test script and design

After consideration of the context of the proposed test, the following observations were made:

1. The proposal (above) for a live video recordings to be played on the computer (while the test taker watches, listens, and responds to prompts or questions) was rejected for two reasons:
    a) Technically, it would be difficult to make changes or edit what has been recorded, if it is a video production. At this early stage, we need to make allowances on how we can develop this test so that after the first version, other versions with different content are possible; an audio recording would be much easier to edit.
    b) The test taker has to be looking at the screen, and listening to the discussion at the same time, and this can affect his/her concentration and thought processes

2. It was decided that a more reasonable version of the test would be as follows: The test taker looks at a screen, hears a voice that gives the introduction and

instructions for the task. He/she then sees two faces on the screen (still pictures), followed by their voices as the discussion begins (audio). As one of the faces turns to the test taker, and a question is presented, the test taker makes his/her response within a given time.

Note: This format is more advantageous as it is possible then to edit the voices or make any other technical changes.

3. The next stage of the study will be developing the computer test as follows:
    a. Record the voice(s) using the existing script for semi-direct 'interactive' task, into a software such as Divace or other compatible voice recording software; after the recordings have been finalized, they are transferred to the computer
    b. Pictures of the speakers are added into the software used for the test; the test is then ready to be trialled (and later edited)
    c. *Trial the test on several ESL/EFL speakers
    d. If it is burnt into a CD, this can be sent to Malaysia for further trials with Malaysian students or any others (before the next Pilot study)

    Note: After each trial with a participant, discuss with the participant the process, and functions that occur or emerge

4. A meeting with associates at CRTEC was conducted for the first viewing of the above test. It was decided that the following changes are to be made:

    a) Rubrics > to include information at start + end of speaking time

    b) Content > to revise as per discussion and noted in sample script copy

c) Differences between the 'new' compared to the 'original' test can be summarized as follows:

	Original test	New test
Turns	Long turns only with better student	same for all test takers
Length	Better students speak more times	same for all test takers
Functions covered	Limited/No interactive functions Limited/No agenda mgt	same for all test takers No agenda management

d) Main advantage of the computer test is it does not suffer from co-construction of discourse which is prevalent/problematic in direct test

e) We need to ensure that functional range of even the best students is not as many as is possible in the computer test

5. The revised version of the computer test :

Pictures ~ (2-4) Speakers talking to each other

Each speaker facing test taker

Speakers talking to each other

Audio recording ~ approximately 6-7 minutes, inclusive of speaking time for test taker

6. On the **functions** of Speaking in the test:

- There may be overlap in the way each function or skill is defined and listed in various sources (Riggenbach (98), Bygate (87), the checklist (O'Sullivan, Saville & Weir 2002), Brown (PhD thesis 2004)

- There is difficulty (as discussed in the references above) in determining whether a function is actually happening, i.e. it is difficult to operationalize some of them, for e.g. 'turn taking' does not appear in Riggenbach's list but in Bygate it appears in 'management of interaction'
- Based on evidences above and that found in the UiTM speaking test where criterion such as 'turn taking' is not included as it is not listed in the syllabus as a micro skill, some items can be omitted from the checklist
- On **implications** for the semi-direct computer test:

  - Since the semi-direct test should be equivalent to the direct test in as many aspects as possible, it is crucial that the functions found in the direct test is included in the semi-direct test

  - The semi-direct test will have the advantage that it can build in some of the functions that do not occur in a direct test; more functions can be elicited

  - In UiTM test, the discourse is co-constructed, as happens in most group discussions of this nature, but the semi-direct test can control this through similar/parallel input for all test takers. This is important as one of the major problems that affect scoring validity; rating is affected when discourse is co-constructed by the speakers, the process and final outcome are unpredictable (see Luoma 2004, Brown 2004)
- The major concern now is the extent to which these functions/skills occur in the course of the test task
- Hence, the new semi-direct test:

  - must meet the criteria on the checklist as much as possible

  - is able to incorporate wider functions than the direct test

- can control for co-construction of discourse aspect

7. Further changes on the CBT after discussions in December 2004:

- ♦ *Thematic link* between task A & B (as in the original direct test

- e.g. Topic: Organizing activities for student visitors

In task A: Talk about 'Places to visit in Malaysia'; include location, cost, entertainment & shopping

In task B: Discuss 'Malaysian way of life'; include food, celebrations/festivities, language and entertainment

- ♦ Planning/preparation time is at the start in 'Introduction' phase, not during the test. This provides an activated *schemata* for a candidate such that during the task, test taker responds at 'real' time; this is more natural
- ♦ Providing some '*structure*' for the dialogue; candidate has points to talk about, and prompts in dialogue from other speakers help too
- ♦ Conclusion is to be decided on one of many items discussed
- ♦ Process of developing the script is *iterative;* the speech by test taker in the dialogue is also iterative

The Computer-delivered speaking test parallels the original (UITM) direct test in many aspects and has the following characteristics:

- Format : 2 tasks Task A Individual presentation
  Task B Interactive task
- Content: Topic(s) selected from past papers of the UiTM direct test

- Thematic link between task A & B
- Preparation times: 1 minute for task A & 1 minute for task B are built into the test. This provides candidates with some schemata/plan for what to say as they are listening to instructions, looking at prompts, etc.

c) Criteria for rating the Speaking Test

Discussions on criteria for rating the speaking test went as follows and decisions were made:

1. Compared rating criteria for different tests such as the TSE, TEEP, IELTS, CPE, and CELS > looks like the IELTS & TSE are the ones most researched & used to various degrees by major exam organizations
2. A table was drawn up (see Table 3.7 below) to compare them & later collapsed to eliminate the ones not useful (e.g. discourse management) and to put together the common ones (e.g. grammar, fluency, coherence)

Figure 3.7 Comparison between rating criteria in test scales

CRITERIA	SCALE				
	IELTS	CPE	CELS	TSE	WEIR (93)
Fluency	√			√	√
Vocabulary	√	√	√	√	√
Coherence	√	√	√	√	
Pronunciation	√	√	√		√
Grammatical accuracy	√	√	√	√	√
Interactional strategies		√	√		√

3. It was decided at this point that the TSE criteria were most useful and closest to the UiTM scale; work commenced from there to determine:

   ~ 2 sets of descriptors for generic criteria for monologue (long-turn, prepared talk) and interactive task (short-turn, spontaneous speech)

   ~ literate vs. less literate criteria (especially for interaction)

   ~ the TOEFL website for latest speaking test TAST & new rating scale, paying attention to descriptions for the less/non-literate criteria for interactive task

4. Comparing the TSE, New TOFLE Scale, Cambridge Common Scale for Speaking; discussions continued with these options:

   ~ For *monologue*> use TSE & new TOEFL speaking scale/descriptors

   * Important decision on the one that is practical for examiners/lecturers or compare the scales with the existing criteria and to be investigated with examiners/lecturers in pilot study Jan-Mar 05

   ~ For *discussion*> see Cambridge common scale for speaking (for all levels of main suite exams) and adjust for UiTM speaking test

5. Use of new *iBT TOEFL Scale* for both Tasks A & B

   After discussions with experts at CLARe & looking at several scales from established exams, a decision was made for the TOEFL scale which is comprehensive in its descriptions (see http://www.ets.org) for details)

   Hence, on the rating criteria:

   a) Using the new TOEFL scale for both tasks A & B because it would be difficult & confusing for raters to switch between rating scales, esp. with the criterion 'grammar'

b) Our test mimics the idea that a three-way conversation has 'chat' (short turns) and 'chunk' (longer turns) in the turns that the participants take (see Eggins & Slade 1997), hence whether for an individual presentation or an interactive task, the criteria for rating can be the same

c) Task A ~ covers a wider range of descriptive, elaborative type functions
Task B ~ cover specific functions (agree/disagree, justify, clarify etc.)

d) Feedback from raters (in MS1) indicated that **existing scale** lacks detail and clear definitions that will enable raters to use consistently; need for a more comprehensive yet practical rating scale

e) The new TOEFL scale, though rather detailed, wordy & can be difficult to use (as it was developed for commercial use), is a good starting point for developing a more balanced scale for UiTM ; it is well-researched, is used for an established exam, has detailed descriptions of characteristics for each band, and is holistic in nature (practicality)

d) Questionnaire for the CB Speaking Test

The next stage is to develop the questionnaires for CBT trials and the following main points needed attention:

a) Purpose: to find out what candidates do at both planning and speaking stage

b) Use of existing questionnaire (from monologue trials) and adjust/reword them for UiTM students; hence, revising from monologue questionnaire + original speaking questionnaires to new questionnaires for computer test:

- Context validity (Your opinion about the test....)
- Theory-based validity (What you did during the test...)

(see Appendix 3.9 for questionnaires)

The 'new' test and the questionnaires were ready to be used in the pilot study.

B. Pilot study (Jan-Mar 2005) in Malaysia: Trials of computer- delivered test (both tasks A & B)

- As the requirements for the computer-based test had been established in Pilot study I with encouraging results, work on developing the complete test which would include task A (individual) and task B (group) was conducted between October and December of 2004 (see section II above for details of the test development)
- However, it is important to note that in spite of the short period of less than four months to develop the complete test, it was essential that we met the deadline of Jan-Mar 2004 so that further data collection work could coincide with the direct test administration in March.
- The computer-based speaking test was thus set up and administered in Malaysia in July through September 2005 as follows:

I. The test was adapted to enable candidates to access it through the web and not have to manage it in its original Power point format (see description of problems with the test in section II above). These enabled each candidate to type in a URL and have immediate access to the speaking test; all other aspects of the test remain the same as the original.

II. Hence, a candidate would go through the following process:

- Listen to and read the instructions and information for the test tasks on the screen

- Given time to prepare during the test (on line preparation, not extra time), and time to speak throughout the test

- Each candidate's speech is recorded into a built in recorder, and each speech segment saved into separate folders; these are then compiled, edited and burnt into CDs for further analysis

- At the end of the test, candidates completed a questionnaire for context validity and theory-based validity

Note: The above procedure in preparing the test involved many hours on the part of the researcher and two technical assistants; however, the outcome was also positive and encouraging as the infrastructure was ready for candidates to be able to access and take part in the test readily.

III. Before the test was administered, a workshop was conducted with members of staff from the Language Centre to discuss issues and obtain feedback related to the speaking test, introduce a new rating scale for speaking, introduce the new computer-based speaking test and engage lecturers, examiners, administrators and experts in rating process using the old and the new scales.

Note: All information and data gathered from the workshop (see Appendix 3.10 for details) were vital and helpful for the study; they would be used in further analyses of data on both the direct and semi-direct computer tests, and in the course of improvising and refining the computer test.

IV. The 'new' computer speaking test was administered between August and September of 2005 as follows:

- 3 groups of 60 students participated in the test in three different sessions in a computer lab; they were given a brief introduction about the test and instructions on how to save their speeches
- Students completed a computer familiarity questionnaire before the test, and a questionnaire on context and theory-based validities after the test, before leaving the premise
- The test takes each candidate up to 20 minutes to complete
- All candidates begin and end the test at the same time

Note: The tests went on quite efficiently with minimal problems and maximum cooperation from all participants and staff members involved.

The following is a detailed account of the Pilot study and all information stated above.

1. Pilot study (conducted Feb-Mar 2005):

A. Preparations were made for the trials at the computer labs on the Shah Alam campus:

- Schedule for the trials were arranged with lecturers involved from the following courses: Pharmacy/PH110; Accountancy/AC110; Applied Science/AS114; Architecture/AP117

- The computer test was inspected and revised by lab technicians; voice recorder built in, instructions revised for students on each power point page, etc.
- Test was installed into computers in Lab 1
- Trials were conducted as follows:

21 Feb: AC110/ 7candidates

22 Feb: AS114/ 8 candidates

AP117/ 7 candidates

PH110/ 18 candidates

Total number of candidates: 40

- After each session, questionnaires (TBV) were distributed & and completed by candidates; the same candidates were also given questionnaires (CV) after they had completed the direct test later

Instructions to candidates

1. Copy all recorded work into a new folder labelled own name
2. SAVE before each recording: in MP3 player, click FILE, type in name, save; then only RECORD doing this each time speech is to be recorded

Problems faced during the trials:

1. Students working at their own pace, hence had control of the slides; the advantage of this is they were able to re-record speech that was unsatisfactory or of poor quality
2. Technical problems in saving, recording, labelling, etc.

3. The test needs further revision/improvement so technical problems can be addressed, & students will not have problems during the test

2. March 2005: A presentation on 'Developing criteria for rating a speaking test' (see Appendix 3E in CD attached) was conducted at the university to begin data collection and to obtain feedback from members of staff who were involved in the test. Preliminary results were as follows:

Participant feedback/comment

i. On existing scale: Examiners/lecturers rarely speak up or inform the Testing Committee of the English department about the use of the scale, vague descriptors, that it is unfriendly and subject to interpretation; it is in need of revision urgently

ii. On new TOEFL scale: Wordy, detailed, but satisfactory & speaking test team needs to consider its possibilities; lecturers/examiners need more time to inspect and familiarize themselves with it

iii. On computer speaking test: Recordings were not satisfactory for use in a workshop; future trials need better quality recordings

iv. Most lecturers are not familiar or knowledgeable on the issues:

- Literate vs. less literate criteria for rating speaking (see Hughes 2002/ 2004; Raof 2002)
- Co-construction of discourse is a major problem in group interaction task (see Brown 2003; Luoma 2004; Weir 2005)
- On computer-based tests, especially in speaking

3. Conclusion from pre-trials and pilot study

- There is support in terms of infrastructure, test participants and members of staff from various institutions in Malaysia which will enable the computer-based test to be used and further developed
- The theoretical and practical research into developing the test structure & design, and the criteria for rating the speaking test (both direct & CB tests) had been conducted to a large extent
- Major concerns raised in the pre-trial and pilot study are found in the table (Figure 3.8) below:

Figure 3.8 Concerns with the computerized speaking test

Test	Concerns
Direct test	Test security
	Equivalent forms in multiple papers
	Co-construction of discourse in group task
	Practicality and uniformity of test administration & management
	Rating & criteria for rating (scoring validity)
Computer test	Mode of delivery/technical issues
	Similarity to 'real' speech event (validity)
	More data needed for theory-based validity

In view of the concerns raised by staff members above, data from students in the pre-trials and pilot study, a workshop was conducted (details below) to gather more data

from members of staff at the university, and to enable them to view and examine the computer- delivered speaking test.

4. May 2005: Workshop on testing of Speaking with participants from the university in Malaysia to gather further data on rating the speaking test
(see Appendix 3.10 for a report + Appendix 3F in CD attached for questionnaire data)

- The main purpose of the workshop was to gather feedback from members of staff (lecturers, examiners, administrators, experts) of the English department on the speaking test, with the intention of introducing to them the computer-delivered speaking test and the new criteria for rating speaking.
- For this purpose, data was collected based on work already completed to date by the researcher and associates at CLARe, and the socio-cognitive framework for validating the speaking test (Weir 2005). According to the framework, the components we look at to ascertain if our test measures the construct it claims to measure are related to contextual features of the test task(s) (context validity), internal processing involved when candidates attempt the task (theory-based validity), and how the test is rated (scoring validity).

The following is an account of activities that took place during the workshop, and information gathered during the workshop, including data obtained from participants who took part in rating exercises for the direct test and the computer-delivered test

The activities:

1. Discussion and brainstorming sessions on the existing test & rating scale
2. Discussion and brainstorming sessions on the computer-based test & the new TOEFL rating scale for speaking (see iBT TOEFL standards for speaking in http://www.toefl.org)
3. Rating the direct test using both the old and new scales
4. Rating the computer-delivered speaking test using the new TOEFL scale
5. Complete questionnaires for both direct and semi-direct tests on context validity, theory-based validity, scoring validity.

The information gathered: Comments and feedback from participants at the workshop were gathered through questionnaires and discussions

<u>On context validity</u>

- Interlocutor factors such as speech rate, accent and acquaintanceship affect UiTM students' performance on the test
- Differences in language ability and degree of interaction for task B are apparent between faculties; students from Business school perform better than those in Applied Science where the examiner gives more help to encourage candidates to speak
- Importance of standardization in administration of test across faculties & in the branch campuses
- Students should be taught how to initiate, maintain and conclude a discussion

On scoring validity:

- Raters are interpreting each criterion differently; this is evident in the data gathered for rating the direct test:

On rating the **direct test** using the **existing scale** > the range in scores given were wide, showing a wide range of ability; we want this gap to be narrowed

On rating the **direct test** using the **new scale** > the range seems to have narrowed; an encouraging indication of the difference between this scale & the old scale

Issues raised on rating:

- Existing rating scale not explicit enough
- Lack of standardization process among raters
- Separate scales needed for oral presentation & discussion
- Analytical scale with clear descriptors and reduced range

Note:

- There appear to be some positive results for context and scoring validity for the computer test at this point, e.g. on scoring validity where upon actual rating of the tasks using the old and new criteria/rating scales, response were positive compared to rating the direct test using the old criteria/scale
- Examiners were exposed to issues surrounding the direct test especially on rating, equivalence of input, standardization, co-construction of discourse

The following is a summary on **rating,** conducted on the direct test and the computer test by workshop participants

a. Rating the direct test using

    a) existing scale

    b) new TOEFL scale

b. Rating the computer test using TOEFL scale

    a) by assigning overall score

    b) by assigning score for each criterion

Description of the outcome of the rating

The following conclusions were made based on the rating exercise of this group of raters who consist of course lecturers, examiners, administrators and subject experts (see Appendix 3G in CD attached for data on ratings)

1. On rating the direct test using the existing scale

    - As a group, the overall range in scoring is large; very few raters had a narrow range in the scores for both tasks across four candidates
    - Some outliers (extreme scores; too low: 4, or too high: 12.5)
    - The objective was too make this gap narrower
    - Further discussion indicated raters had their own interpretations of what each criterion (e.g. CA) meant

2. On rating the direct test using the new TOEFL scale

    - A quick analysis of the data showed raters are rating much closer on the new scale in spite of absence of training and exposure to standardization tapes, than on a scale they had been using for some time (the old scale)

3. On rating the computer test using TOEFL scale

    - by assigning overall score

- by assigning individual score for each criterion

- Overall range is narrower than previously seen; very few inconsistencies in scoring
- On correlations : All raters are rating four candidates on the same test using different rating scales > between rating the speaking test using the UITM scale and rating the speaking test using the TOEFL scale:
- Correlation indexes are low:

  For candidate 1 = 0.30

  2 = 0.54

  3 = 0.49

  4 = 0.26

- No significant correlation or relationship in rating the candidates using the different scales

  Probable reasons for these correlations:

  - Rating scales are different in terms of clarity and specificity of descriptors
  - Range is wider when the test was rated using the UiTM scale
  - Feedback from participants showed that they would prefer the UiTM scale to be revised; each criterion needs to have clearer and more specific descriptions
  - As a group, raters seemed comfortable using the new TOEFL scale, in spite of lack of information/familiarity and training

4. Feedback on the *Computer test*

a. Features and potential benefits of the computer test:

i) Equivalence of input for all candidates in terms of instructions (speed, accent, linguistic & functional features are the same for all); this addresses the issue of "co-construction of discourse" which affects performance and rating

ii) Equal time to prepare and to respond

iii) Reliability of scoring (e.g. pass/fail cases can be easily identified)

iv) Interactional feature was evident between some candidates & input received; it seemed that some candidates' speech improved as they progressed through the task, especially in task B where there were four prompts for them to respond to, i.e. performance is incremental

v) There is a permanent record; tests can be re-played & re-evaluated

vi) There is potential for standardization across the board, especially across campuses

vii) Task seems less stressful for candidates, especially with absence of examiner & other speakers

viii) Suitable for computer generation

ix) No need for interlocutor training

b. Potential problems of the computer test

i) Effects on teaching; major change in teaching paradigm

ii) Infrastructure and cost

iii) Test security

CONCLUSION

1. The direct test

On context validity:

Data gathered from the questionnaires indicated that the areas of concern are weighting for each task, known criteria for rating, nature of information, linguistic and interlocutor variables.

On theory-based validity:

Lecturers/examiners did not agree that students were able to check and make adjustments to the structures they used in their speech, monitor their speech in relation to other speakers, and manage interactional functions during the group discussion task.

On *scoring validity:*

Response from questionnaires and discussions indicated that a major source of disagreement occurred for scoring in terms of clarity of criteria for rating, rater training & standardization, rating procedures such as consistency in rating and moderation, and grading and awarding procedures.

2. The computer test

On context validity:

Response here were favourable for all aspects of the test task

On scoring validity:

Response were positive for items related to criteria and scale for rating, and potential for intra as well as inter-rater reliability

Overall, there seemed to be some positive feedback for the computer-based test compared to the direct test; though more data is needed to further support these claims. But for this group of participants who were exposed to both context as well scoring validity aspects of the tests, the impact on them was coherent and positive.

There is yet a lot of work to be done and research to be conducted on the speaking test. Validation is the beginning of this process as we attempt to improvise our practices and the quality of our test. Validity is not a property of the test instrument itself; it is what provides tests with quality as we concern ourselves with the soundness of the interpretations and uses we make of our test results. Using a socio-cognitive framework (Weir 2005) will enable the process to be more systematic and meaningful, especially since the relationship between the test task (context validity), the test taker (theory-based validity) and how we rate the performance (scoring validity) is crucial in determining construct validity. Further investigation is needed in areas of interaction in the speaking test (group work), on refining the computer test & gathering more data on it, and validation of the computer test.

Thus far, the theoretical design of the study as it was first conceived and the bulk of the information above relating to the development of the research instruments for the study, were presented in detail. As reported above, the questionnaire and interview guidelines for speaking (relevant for both the direct test and computer test) were developed over several months and were tried out and piloted many times before they could be used in the main studies (findings found in Chapters 4 & 5). Moreover, additional data for use in

the main studies were gathered through a series of presentations and a workshop on the testing of speaking.

The next section provides details of the participants involved in the study and the locations where various stages of the study take place. This is then followed by an operational model of the study, which was conceived following the outcome of the development and trials/pilot studies of the research instruments; this is a model of how the study was essentially conducted.

SECTION 111: Description of participants and locations

Participants

Participants for this study (both Main study 1 and Main study 2) will consist of those who are involved directly, and indirectly, in the speaking test, and will be described below for each stage of the study:

a) **Students** (full-time and part-time) from Mara University of Technology who are in at least semester 3 of their studies. A representative sample will be selected based on their groupings in each of the 6 academic programmes (15 faculties).

Stage 1: Approximately 80 Participants will take part in a questionnaire survey

Stage 2: Approximately 150 Participants will take part in a questionnaire survey

30 Participants will take part in an informal interview session for data on

Stage 3: Approximately 500 Participants will take part in a questionnaire survey

50 Participants will take part in an informal interview session for data on

Stage 4: Approximately 100 Participants will take part in the trial stage of the new semi direct computer-based test and respond to a questionnaire survey related to the test.

Stage 5: Approximately 100 Participants will take part in the new semi direct computer-based speaking test and respond to a questionnaire survey related to the test.

50 Participants will take part in an informal interview session for data on

b) **Lecturers and examiners** from the Language Centre and lecturers from other faculties.

Stage 2: Lecturers and examiners from the Language Centre will take part in questionnaire survey and an informal interview session on their views of the existing test in the pilot stage of the research

Stage 3: Lecturers and examiners from the Language Centre will take part in questionnaire survey and an informal interview session on their views of the existing test in Main study 1

Stage 4: Lecturers & examiners from the Language Centre will be consulted at the design and development stages of the new semi-direct computer-based test

Stage 5: Lecturers and examiners from the Language Centre and lecturers from other faculties will be briefed on the final results of the study, and presented with recommendations for a new test of spoken language

c) **Administrators**: Deans/Heads, Deputy deans/heads (administration & academic), Course tutors/coordinators/representatives from the Language centre and EDC

Stage 2: Administrative and academic representatives from the Language Centre, and will take part in a questionnaire survey, and informal interview session in the pilot study.

Stage 3: Administrative and academic representatives from the Language Centre, and will take part in a questionnaire survey, and informal interview session in Main study 1.

Stage 4: Deans/Heads, Deputy deans/heads (administration & academic), Course tutors/coordinators/representatives from the Language centre will be consulted at the design and development stages of the new semi-

direct computer-based test

Stage 5: Deans/Heads, Deputy deans/heads (administration & academic), Course tutors/ coordinators/representatives from the Language centre and EDC will be briefed on the final results of the study, and presented with recommendations for a new test of spoken language.

d) **Experts** in the field of education and language testing from the Malaysian Examination Board and Department of Education

Stage 2: Representatives from the Examination Board and the Department of Education will take part in an interview session to gather their views on the existing MUET speaking test

Stage 3: Representatives from the Examination Board and other subject specialists will be given questionnaires and take part in interview sessions for Main study 1.

Stage 4: Representatives from the Examination Board and the Department of Education will be consulted at the design and development stages of the new semi-direct computer-based test

Stage 5: Representatives from the Examination Board and the Department of Education will be briefed on the final results of the study, and presented with recommendations for a new test of spoken language

Location/context

- Piloting of framework instruments and procedures in stage 2 took place in the UiTM campus in Malaysia

- Evaluating the existing speaking test took place in the UiTM campus in Malaysia and CRTEC, University of Surrey Roehampton

- Piloting the new semi-direct speaking test took place in the UiTM campus in Malaysia

- Evaluating the new semi-direct speaking test took place in the UiTM campus in Malaysia and the CRTEC, University of Surrey Roehampton

- Main trial of the new test took place in the UiTM main campus in Malaysia

- Final review, revision, and analysis of the study were conducted at CLARe Roehampton University

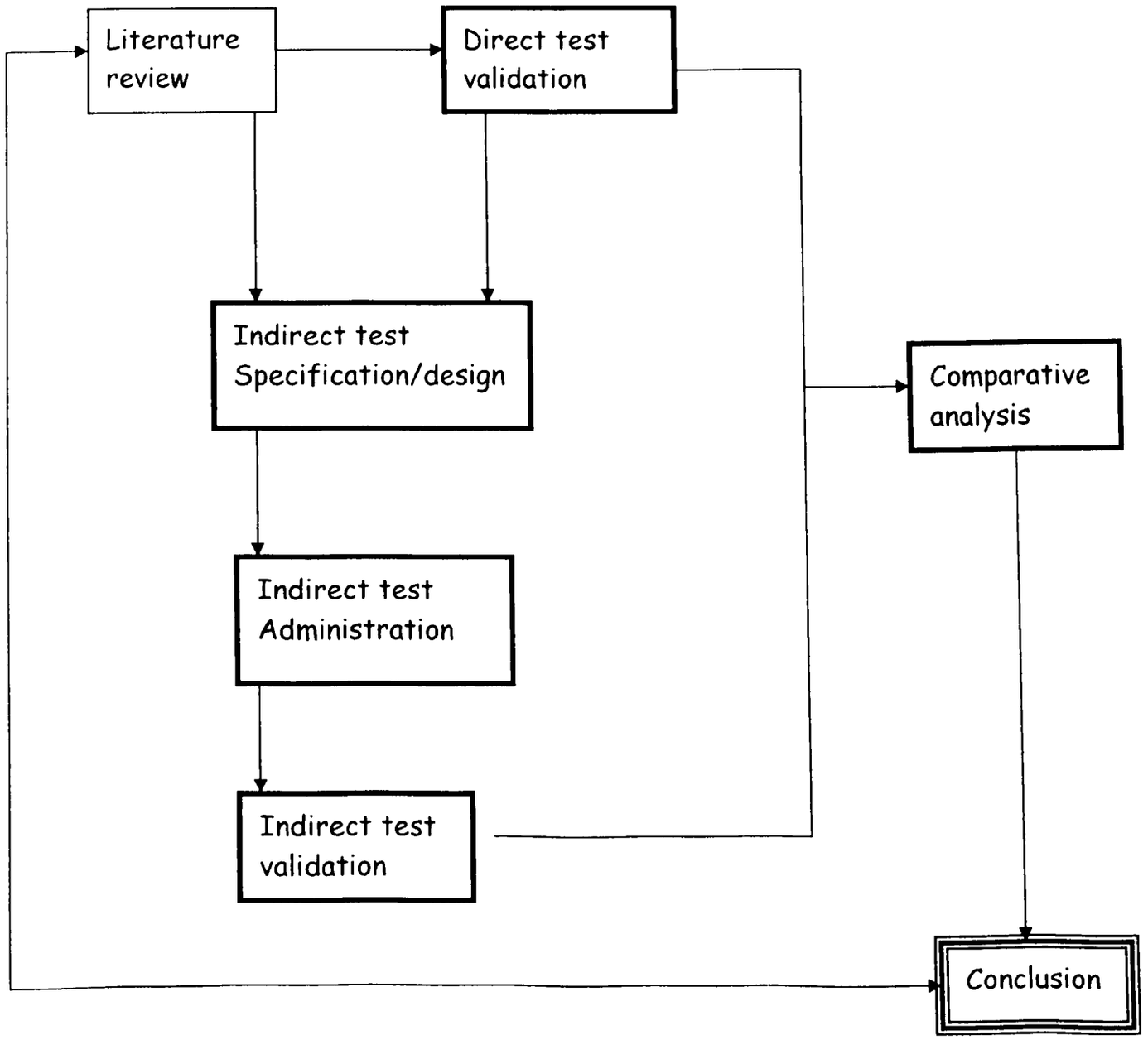
SECTION 1V: Model of the research plan

A model was developed (see Figure 3.9 below) to illustrate how the research would be conducted; this is the operational model of the study, as opposed to the design in section I, which was a theoretical model. Sections II and III presented detailed accounts of the research instruments, including the pilot studies which were conducted as part of the process of trying out and refining the instruments.

The model illustrates different stages of the study, which will be outlined and discussed individually in respective chapters. There are five stages and each stage will be presented separately and adjustments will be made to each one according to how the studies progress in terms of the types of data and how much data are collected, and the analyses of the findings.

- Stage 1: Direct test validation includes Main Study 1 which will be presented in chapter 4
- Stage 2: Indirect test validation includes Main Study 2 which will be presented in chapter 5
- Stage 3: Indirect test administration includes the pre-trial and pilot study described above, and the administration of the final version of the test in Main study 2 (described in chapter 5)
- Stage 4: Indirect test validation provides details of the findings from the pre-trial study, the pilot study and Main study 2; these will be presented in chapter 5
- Stage 5: This final stage of the study involves a comparison between the findings from Main study 1 and Main study 2; these will be presented in chapter 6

Figure 3.9 Operational Model of the Research Design



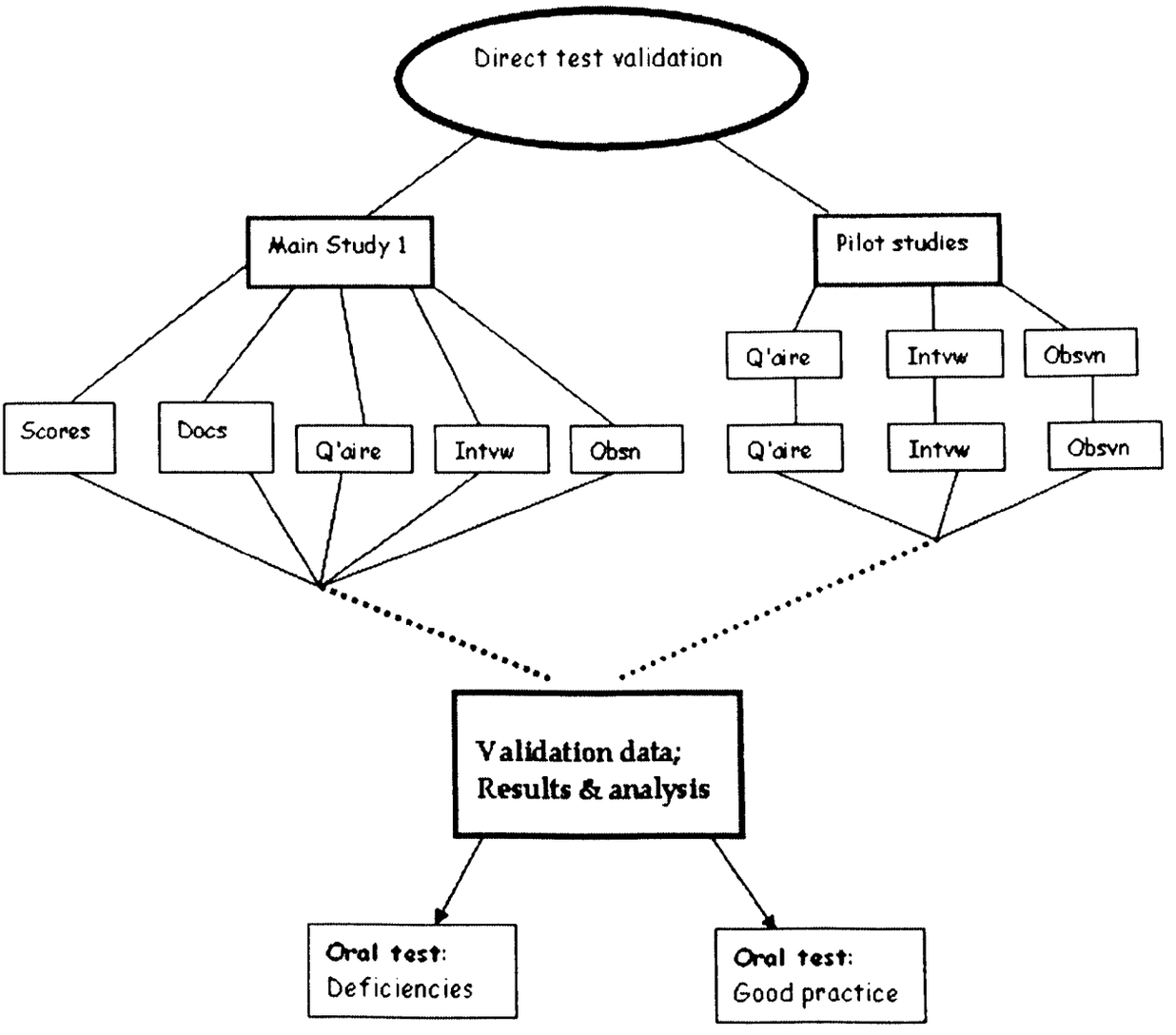
Chapter 4: DIRECT TEST VALIDATION

4.1 INTRODUCTION

This stage of the study involves the validation of the existing speaking test in UiTM Malaysia. It began with a review of the literature on validation and test validity, testing of speaking, and testing using the computer, all of which are found in chapter 2.

Before the test could be validated, two pilot studies were conducted to tryout the instruments (questionnaire, interview) and check other arrangements such as venue and participants for the study. The validation study or Main Study 1 was conducted in January-March 2004 and the main sources for validity evidence were gathered from the questionnaires, interviews, observations, course and test documents and test scores. These data were organized and analyzed and the findings are presented in this chapter in terms of the framework for test validation. The aim of the validation study was described in chapter 3, i.e. to establish the strengths and weaknesses of the test in terms of its validity and reliability as a test of a learner's ability to speak in the target language (English) in academic and social settings, through a presentation and group discussion. All this is illustrated in Figure 4.1 below which is the design of the study.

Figure 4.1 Design of the study



This chapter presents a report on data collected in Malaysia for Main Study 1 (Jan-Mar 2004) of stage 1 (direct test validation) of the research. In chapter 3, we saw how the research had been designed and how the research instruments were developed in line with the research questions and different phases (chapter 3; section I) and stages of the research (chapter 3; section V). In this chapter, the findings from Main Study 1 will be presented in the tables below.

This stage of the research attempts to answer the first research question of the study, which is:

1. To what extent is a face-to-face speaking test valid in terms of:

a) context validity

b) theory-based validity

c) scoring validity

The speaking test in question has been in existence for the past five years at Mara University of Technology Malaysia, and it is a part of the final exam in an English Proficiency course for students who are pursuing a diploma in various fields at the university. It was developed based on the national test MUET (see chapter 1: MUET) as a preparation for students to take the national test, a two-pronged strategy which apparently would help students with their proficiency in the language, and indirectly prepare them for the test. In spite of its importance, no validation studies have been conducted or undertaken to explore the issues of reliability and validity of the test. In the Malaysian context, very little work has been done on the validation of direct or indirect tests of speaking (see chapter 1 for details).

Literature on the testing of speaking also indicates that validation studies in this area of language testing are scarce. This research, therefore, is an attempt at validating the UiTM speaking test, and this was done through a process which operationalised a framework for validating a speaking test (Weir 2005). The outcome of this validation

exercise will in turn enable us to explore further the possibilities of a semi-direct, computer test, which we hope would be able to address problems faced by a direct speaking test, such as the one used at the university.

All the data for Main Study 1 were gathered through interview sessions with students, and members of staff (lecturers/examiners, administrators and experts) at the university, a questionnaire survey administered to the same participants, a collection of test documentations, and several observations of the test by the researcher. All this data were organized and analysed as follows:

- Interview data (qualitative data) were organised/analysed using *Hyper Research* software for qualitative data analysis
- Questionnaire data (quantitative data) were organised/analysed using *SPSS* software for quantitative data analysis
- All this evidence, plus other sources of data, such as documentations from the test and observations, were entered into a matrix/table which shows the source of validity evidence and the validity components (see Appendix 3A in CD)

4.2 THE SPEAKING TEST

The speaking test is briefly described in Figure 4.2 below:

Figure 4.2 Features of the direct speaking test

Test type	Objective	Format	Test input	Conduct
Direct/ face to face test	To test students ability to speak in the target language in an academic and/or social setting	• Four candidates (A, B, C, D), and two examiners (usually class lecturer + 2nd examiner) present in a test session •Two tasks: Individual presentation + Group discussion	• Question paper (written) • (Oral) instructions on test conduct given to candidates when they enter the room	• Each candidate participates in two tasks: individual presentation (task A) followed by group discussion (task B) •Preparation time: 2 minutes for both tasks •Presentation time: Task A ~ 2 mins Task B ~ 10 mins • Both examiners listen to and rate presentations

Note: There is a thematic link between topic in task A & task B; what each candidate presents in task A are important considerations for the discussion that follows in task B.

4.3 FINDINGS MAIN STUDY 1

(see Appendix 3M in CD I attached)

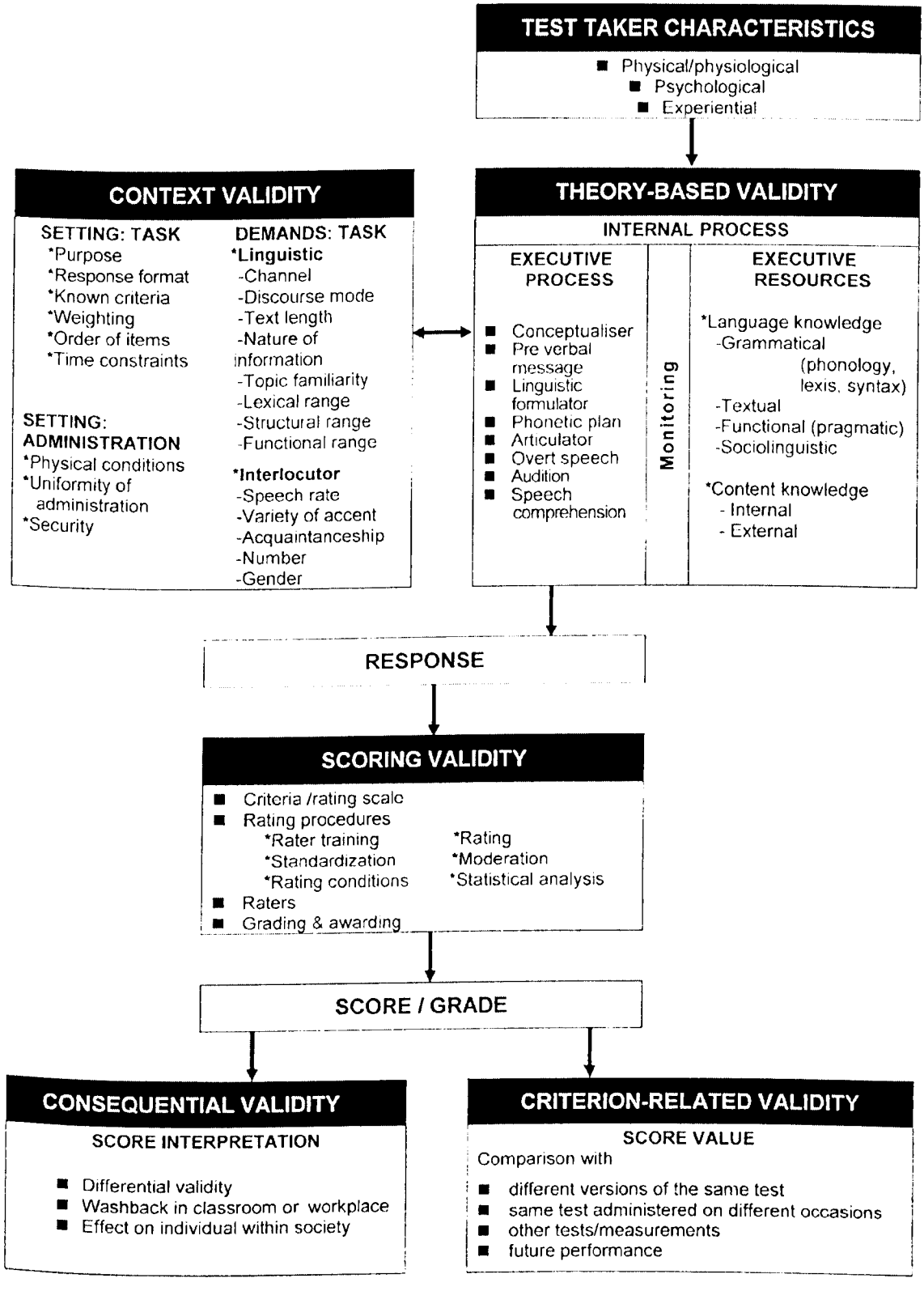
The socio-cognitive framework for validating a speaking test (Weir 2005) was used from the start of the study to develop the research instruments, and data gathered were also reported according to the validity components of the framework. It consists of five components, and this study addresses the three components which lie at the heart of construct validity, i.e. context, theory-based aspects of validities and scoring criteria (see chapter 1 for further explanation on why the study focused on the three components). How each element, such as the purpose of the test, is expressed to the test taker will

influence how he/she performs the task and succeeds. The scoring criteria, which are developed based on the test context, would directly affect test performance, especially when they are made known to the candidates. Any evidence of these validity components and how they interact will give us a better picture of whether the test is measuring the ability it claims to measure.

At this point, we note again that these components of validity are presented in the Figure below (Figure 4.3). There is a temporal sequence as well as a conceptual representation of the components, i.e. the timeline in terms of when each component happens (e.g. context validity before scoring validity), and the relationship between them (e.g. test context affects what and how the test taker does to perform the task, i.e. theory-based validity). In essence, a person's ability to speak does change, depending not only upon whom he/she speaks to, where and what about, but under what conditions (Fulcher 2003).

Thus, data gathered are presented according to the connections and relationships between validity elements.

Figure 4.3 Socio-cognitive framework for validating a speaking test (Weir 2005)

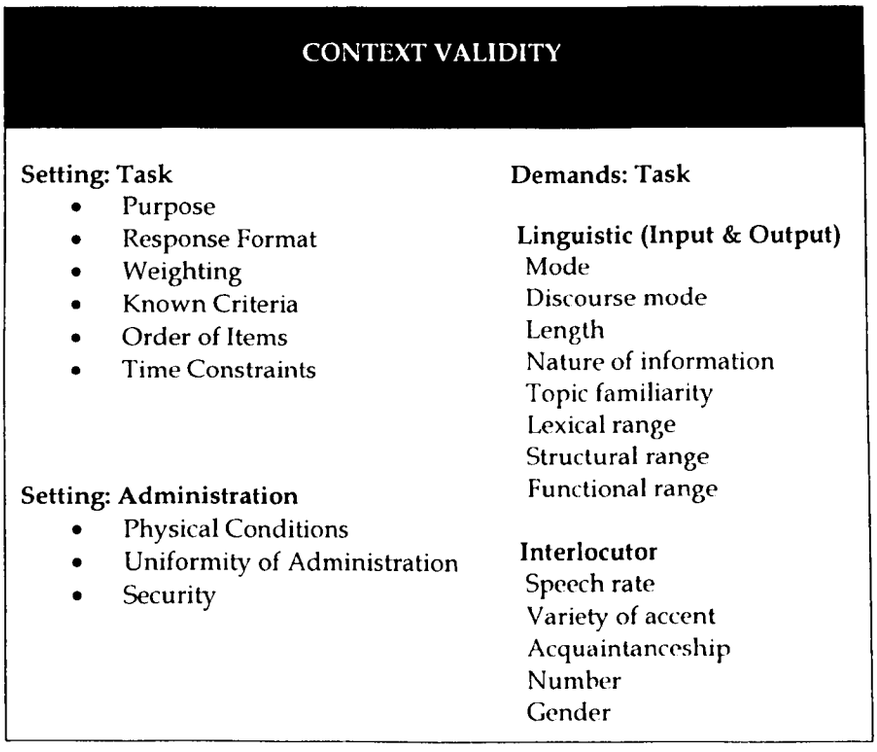


4.3.1 CONTEXT VALIDITY

The context of the test task is paramount; operations and conditions of the test need to be spelled out clearly in the test specifications so candidates are clear about what they have to do and how. Context validity is the 'what' and 'how' of the test and is divided into task setting or the more discernible features such as time constraints, test administration, and task demands which include linguistic (input and output) and interlocutor variables. (see Figure 4.4 below)

The aim here is to ascertain if data gathered in this section show that the test fulfils context validity in terms of its elements; more importantly, that the test had been developed based on sound theoretical model and expert judgment on the assessment of spoken language. If it were, our findings below would reflect or reveal this, and the test is said to be valid for the construct it claims to measure.

Figure 4.4 Aspects of Context Validity for Speaking (from Weir, 2004)



The following tables contain the findings of the study according to each validity component, e.g. context validity; each element of the validity component, e.g. test purpose; instruments used to gather the data e.g. interview; and the informants for this data e.g. students. Each validity element has three tables: data from participants*, document analysis, and observation. Brief descriptions of the elements are also presented, followed by a commentary on the finding(s) that show any significance to the research.

* Staff (lecturers, examiners, administrators, experts) were listed separately for identification, and their findings were combined as they represented the same test and were given the same questionnaire & interview questions.

4.3.1 a) Task Setting

This refers to the settings under which the test activity takes place, i.e. the purpose of the tasks, response format, known criteria, weighting of each task/section, order of items, and time allocated for the tasks. These elements should be explicitly expressed in the test so candidates will be able to execute their internal processing and strategies effectively. Hence, in this section we aim to look for evidence of these validity elements through data gathered from the various participants; their responses should reveal the existence or absence of the elements which in total make the test valid in terms of its task setting.

Purpose refers to the aim or requirements of the task.

Test takers should be given a clear idea in the rubric of what the requirements of the task are so that they can choose the most appropriate strategies and determine what information to activate in order to complete the task. Having a clear purpose will facilitate *goal-setting* and *monitoring*, two strategies of cognitive processing that candidates would utilize to attempt the task. In the UiTM test, the clarity of purpose is requisite for both tasks A & B, which are separate tasks but linked by a common theme.

Table 4.1a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	85% agreed purpose for both tasks were clear
Questionnaire	Lecturer/Examiner (N= 46)	100% agreed purpose for both tasks were clear
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Purpose clear; no confusion
Interview	Lecturer/Examiner (N = 10)	Purpose is clear for candidates
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	

Commentary

It appears the *purpose* of the tasks were explicit & made clear to candidates in the question paper, i.e. the instructions on what a candidate has to do in preparation and presentation for each task, are clear. This includes sequence of activities, time constraint for each activity, available input/content to be processed, and the speech functions to be demonstrated. The fact that the administrators and lecturers agreed that the purpose is clear is not really a surprise, in that they are involved with the test in different capacities. While the same might be said of the experts, they should be seen as offering an impartial perspective - though their past experience of this type of test will undoubtedly have contributed to their observations.

In general, we can conclude from these results that there is no problem with the purpose of the test tasks.

Table 4.1b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Purpose stated
Question paper	Purpose for both tasks stated
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	Purpose for both tasks stated

Commentary

Except for the score sheets/scoring guide (scale), all documents stated the test purpose for both tasks: in task A candidates are to present facts/ideas in support of a topic, in task B candidates discuss four of these facts/ideas and conclude by selecting one. This may not be a serious issue, though it would clearly be of benefit to the examiners if the purpose is clarified in the scoring guide - though it should be remembered that the examiners are in fact also the teachers, who unanimously indicated their satisfaction that the purpose of the task was made clear. (see Table 4.1a above).

Table 4.1c

Observation	*Findings*
Researcher	Student reactions and responses during the tests indicated a high level of understanding of task requirements

Commentary

It was observed that for almost 90% of the time, students did not require help from the interlocutor/examiner present for clarification of what was required of them in the

tasks. The results of these observations seem to support the findings from the other evidence reported in Tables 4.1a and 4.1b (above).

In any test, a clear purpose is paramount for candidates to perform the task effectively (see Evans 1988, Moore & Morton 1999, Weigle 2002 on purpose for writing). In a spoken language test, a clear purpose will facilitate the focus, direction and outcome of the interaction for the interlocutors. Though authenticity is sometimes difficult to achieve in most tests, (Kormos 1999, Stansfield & Wu 2001), tasks need to emphasize a realistic and needs-based purpose for the speakers.

The speaking test above may not be authentic but the purpose, i.e. to speak in the target language in a setting familiar to students, seems clear.

Response format refers to the technique required of the test taker in attempting the task, such as short answer questions in reading and writing, which is different from multiple-choice questions in the same tests.

The choice of response format in a test can seriously affect cognitive processing; in a speaking test this could include a narrative on a picture sequence, interaction between student and examiner (free or controlled interview), interaction between peers, and monologue tasks. In the UiTM test, the tasks involve an individual presentation to a small audience (examiner/rater and three other speakers) followed by a group discussion between the four candidates.

Table 4.2a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	80% agreed both task formats appropriate
Questionnaire	Lecturer/Examiner (N= 46)	Approx 80% agreed both tasks reflect student abilities to communicate in academic context; 55% agreed both tasks reflect student abilities to communicate in social setting
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Both tasks format appropriate
Interview	Lecturer/Examiner (N = 10)	Response formats are appropriate for the test
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	Good format as you can judge if candidate can present individually as well as interact with other speakers

Commentary

Most informants seem to agree that the individual presentation and group discussion are appropriate response formats, even though only half of the staff respondents agreed that both tasks were good tests of a student's ability to communicate in a social context, i.e. they are better tests for academic contexts. This low percentage of agreement is a little disturbing since this was listed as one of the objectives in the course syllabus. Further data from the interviews showed that those who disagree that the discussion is an indicator of students' abilities to communicate in a social context cited reasons such as lack of practice on this format, time given to candidates to demonstrate this ability, and topics that do not encourage this ability, instead the difficult topics can stifle a student's ability. They agree however, as evident in the interview data, that these formats are suitable because of the thematic link between them; what a candidate

presents in task A could affect the outcome of the discussion in task B because the topics are related, and candidates are able to demonstrate speaking ability in two different context.

It is evident from the above data that respondents agree with the formats but not for the reasons they were included in the test. It is encouraging to know that some members of staff are able to link one element to several others, i.e. that some elements in the task are inter-related and can affect performance.

Table 4.2b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	Described in terms of tasks
Question paper	Stated in instructions
Score sheet, scoring guide/rating criteria	Information on rating task A followed by task B
Instructions/guidelines	Described in terms of tasks

Commentary

Response format appears in all the documents. The score sheets, scoring guide/rating criteria contain descriptions of how the test is conducted and rated, and the criteria for rating them according to the two tasks, although the relationship between test tasks and criteria selected for rating them is not immediately clear in the documents above (see Weir '05, chapter 9 for the importance of this). Data from staff questionnaire and interview (later in section Scoring Validity) on whether these factors were taken into consideration when the test was developed indicated that the relationship between them was not made clear.

Table 4.2c

Observation	*Findings*
Researcher	Students were very familiar with both formats

Commentary

Familiarity with response format is important as it results in a positive outcome, especially for these test takers who have had practice in both response formats in the classroom. The fact that they are familiar however, does not guarantee good performance during the test, but it does for the most part provide students with a better idea of what to expect in the test.

It was evident during the observations that students were able to conduct themselves efficiently in the test because of this knowledge. However, because the format we choose affects a candidate's cognitive processing of the task, we need to select task formats carefully so they do not adversely affect processing that we would want to occur to answer the tasks we set (Alderson et al 1995, Hughes 2003, Weir 2005). In the speaking test above, students' familiarity with test formats may not necessarily result in a demonstration of linguistic and content knowledge demanded by the test (see tables 4.18a ,4.21a-4.24a below for data on language knowledge under theory-based validity). There is no evidence from the documents to show how the format individual presentation followed by discussion was decided upon.

Weighting refers to the assignment of a different number of maximum points to a test

item, task or component in order to change its relative contribution in relation to the other parts of the same test (Weir 2005).

If different parts of a test are weighted differently, then the timing or marks to be awarded should reflect this; the test taker needs to be informed so that they can allocate their time accordingly in the goal-setting phase of processing. In the UiTM test, the tasks are given equal weighting, 15 marks each.

Table 4.3a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	60% disagreed on equal weighting
Questionnaire	Lecturer/Examiner (N= 46)	60% agreed both tasks should have equal weighting
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	No data
Interview	Lecturer/Examiner (N = 10)	Confirms with response in questionnaire
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	

Commentary

Both groups have contrasting views about weighting of the test, which at present is equal for both tasks. Unfortunately there were no comments from the students on this in the interviews, seeing from the questionnaire that they would like each task to have different weightings. One suspects from talking about other aspects of the test such as

time constraints, linguistic demands and topic familiarity that most students do find the tasks of different difficulty levels, hence why they should not be treated equally in terms of marks and weights.

Staff members' agreement to the equal weighting is not surprising because the test was designed according to the MUET in which the speaking tasks have equal marks and weights. The administrators especially, would expect this, as with other aspects of the test, to comply with the MUET specifications, although the justification for it is not available.

Table 4.3b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	Stated here
Question paper	N/A
Score sheet, scoring guide/rating criteria	Stated here
Instructions/guidelines	N/A

Commentary

Weighting for each task is stated in all documents except the question paper. At the point of doing the test, test takers need to be reminded of how each task is weighted to facilitate processing. It is also not surprising that this information is not clearly stated in the question papers because all lecturers who teach the course are expected to inform and remind their students about it.

Table 4.3c

Observation	*Findings*
Researcher	During test sessions students were not reminded of or informed of the weighting for each task

Commentary

As with all other participant data above, students need to be aware of the marks they can attain for each task; this element is not evident in some vital parts of the speaking test. When they are not aware of or perhaps do not remember this information during the test, it could affect their preparation for the task and their performance. It seems from this that there are difficulties in relying on teachers to take the responsibility of informing the students about weighting; it should be done at the point of contact, i.e. the exam.

This is a crucial point seeing that candidates' time and attention on monitoring their output are directly affected by the weighting of items in a test. Again, there is no evidence in the documents in this test to indicate why and how the tasks were weighted.

Known rating criteria Students should have a clear idea of how they will be judged and the criteria used for judging them well in advance of the test.

Information about how a performance will be scored and the criteria used for scoring have to be made explicit to students as well as markers. As with other elements in context validity, having this knowledge will facilitate a test taker's cognitive processing in terms of monitoring their speech during the performance.

Table 4.4a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	65% agreed criteria clear in test instructions
Questionnaire	Lecturer/Examiner (N= 46)	72&% disagreed criteria were clear
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Many do not know details of rating criteria; some informed of what they are but no explanations
Interview	Lecturer/Examiner (N = 10)	Students are informed in the classroom, not during the test
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	

Commentary

There seemed to be some confusion about this item in the questionnaire because upon checking the question papers (see table below), the researcher found no mention of criteria for rating the tasks in them at all. Students' response might be because they were informed at some point during class practices or lectures. Data from the interviews however, clarified the fact that many of them did not know or were not informed of how they are rated for the tasks.

Staff interview data confirmed that the lecturers teaching the course are expected to inform and explain to their students about this. Unfortunately, some students were not informed or were informed in general terms only, i.e. what the three criteria are but not what they mean and how they are each weighted. Again, there seems to be difficulties in relying on teachers to take the responsibility of informing the students about the rating criteria, which themselves are not equally weighted (see Weir 1983a, 1988c on differential weighting for the TEAP).

Table 4.4b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Not stated here
Question paper	N/A
Score sheet, scoring guide/rating criteria	Includes criteria, descriptors and a scale on which they are distributed, and how scores are recorded
Instructions/guidelines	N/A

Commentary

Seeing that the documents related to the rating process (score sheets, scoring guide, rating scale) are given to lecturers/examiners a few days before the test is administered, it is not surprising that we have evidence such as the ones in Table 4.4a (above). Since the information is not provided in the course syllabus and test specifications, lecturers would not have informed their students about the rating criteria, rating scale and so on well in advance of the test. More importantly, the information is not provided in the question paper and hence, students are disadvantaged because they did not have this information to consider when preparing for the tasks. For example, if they had known that language ability carries more marks than communicative ability, they might monitor their vocabulary and sentence structures more carefully during the presentations.

Table 4.4c

Observation	*Findings*
Researcher	During test sessions students were not reminded of or informed about how they will be rated

Commentary

As stated above, it is unfortunate that test takers are not aware of how they will be rated for their performance in the test because having this knowledge could facilitate goal-setting, planning during preparation time, and monitoring during the presentations. In general, information related to how students are rated in the test is not clearly indicated in the question paper and to students. This affects the context validity of the test where this element is concerned. In addition, data from scoring validity (see table 4.25a below) show serious inconsistencies and problems in rating the speaking test, including the use of a scale and a set of criteria which are vague and lack credibility. All these point to the important consideration of advising students on how they are rated in the test; when examiners and raters are unclear of their tasks and the rating scale, students are affected in their processing of the various criteria required of them.

Order of items refers to the order in which the questions/tasks appear in the question paper. Whether the order of items in the test is justifiable is the main concern here. In the existing speaking test, the individual presentation is followed by the group discussion; this will influence how students process their thoughts in terms of planning and monitoring, especially because of a thematic link between the tasks, information presented in task A will affect the outcome of the discussion in task B.

Table 4.5a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	84% agreed the order is appropriate
Questionnaire	Lecturer/Examiner (N= 46)	91% agreed the order is appropriate
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	The order is appropriate because in the test task A influences what is said in task B
Interview	Lecturer/Examiner (N = 10)	Discussed under 'response format' above
Interview	Administrator (N = 6)	
Interview	Expert(N = 5)	

Commentary

The degree of agreement about task A followed by task B is high between the two different groups of participants; this is not surprising, especially from teachers and administrators because they had been managing the test in this order for six years. It is interesting to note that examiners seem to think that this order helps the weaker students who would take note of language use and content of other speakers, which in turn helps them prepare and perform in task B. Interview data from students echoed this point, while more of them talked about the thematic link between the tasks; what they present in task A could be integrated in the discussion in task B, in fact, it usually is.

There seems to be confusion between the order of the tasks being appropriate because it serves the purpose of the whole test and because of the thematic link between them. While most students think the order is appropriate, they still find making the

connection between points raised in task A and the discussion in task B rather challenging.

Table 4.5b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	Stated here
Question paper	Evident here
Score sheet, scoring guide/rating criteria	Evident in score sheet
Instructions/guidelines	Stated here

Commentary

The order of the tasks is evident in all documents. The syllabus provides a logical reason for this order in relation to the objectives of the speaking component (see chapter 3, section II: Instruments for Main Study 1), while the instructions/ guidelines for administering and marking the test explains this in terms of the thematic link of the topics for presentation.

In general, it would be beneficial for all parties concerned to have a clear and detailed explanation and justification of this information in one document, which students and lecturers/examiners can access when required (see Urquhart & Weir 1998, Kintsch 1998 on order of tasks for reading; Buck 2001 for listening). This is especially crucial for this test; the order was determined because points raised in task A do influence the outcome of the discussion in task B. What should be emphasized is not just A followed by B, but why and how this can be achieved by the student satisfactorily.

The data above (from Table 4.9a) reflect this clearly, i.e. confusion in the reasons for this order of task A followed by task B, is the reason logical, purposeful or the thematic link

between them. Clear information about whether the tasks are so ordered due to logical or affective reasons would help candidates in terms of planning for the tasks and monitoring their own as well as other speakers' speech.

Table 4.5c

Observation	*Findings*
Researcher	The order is followed in every test session; students are familiar with this format

Commentary

It is not surprising that the students are very familiar and seem comfortable with the order of the tasks because they feel they have the required knowledge to respond well and have had practice in their English classes.

Hence, while the order of the tasks is not a problem for the students, administrators and teachers need more information and clear reasons for the order of tasks, and how this affects students during the test.

Time constraint refers to the amount of time given to candidates to complete the the test. This includes time for each section and part of the test. The amount of time allocated for each section should reflect the importance of this element in the part of the course and the domain being tested. Time constraint is an important element which will affect how candidates approach the test for goal setting, planning & monitoring. The time they are to spend for each section needs to be made explicit in the test paper and examiners should encourage candidates to comply with them.

Table 4.6a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	53% agreed 2 minutes sufficient to **present** task A; 77% agreed 10 minutes sufficient to present task B
Questionnaire	Lecturer/Examiner (N= 46)	36% agreed 2 minutes sufficient to **present** task A; 77% agreed 10 minutes sufficient to present task B
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	2 minutes is insufficient time to **prepare** task A; 2 minutes is insufficient time to **present** task A
Interview	Lecturer/Examiner (N = 10)	2 minutes is insufficient time to **prepare** task A; 2 minutes is insufficient time to **present** task A
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	

Commentary

Students and most members of staff agreed that the preparation and presentation time for task A is insufficient. Data from student interviews showed that they were under pressure to read, think critically & sometimes write their points down in a short time; during the presentations, many students were not able to express their ideas clearly and complete the presentation. Others commented on how time adds to the anxiety and other pressures felt by them already. Lecturers expressed their concerns regarding the objective of the test, which require students to demonstrate not just language ability, but a high level of critical thinking and organizing of ideas, all this to be completed within a very short time.

In general, time constraint is a major problem for many candidates in this speaking test.

Table 4.6b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Stated in test specifications
Question paper	Stated here
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	Stated here

Commentary

All documents, except the score sheets and scoring guide/rating criteria, provide information on times for preparation and presentation for both tasks A & B. Both students and members of staff are thus fully aware of the time constraints imposed in the speaking test. This is vital considering the facts that the two tasks are conducted consecutively, content from task A affects the outcome of task B, and in task B, students are involved in an interaction which is limited in terms of content, yet open-ended in how the candidates steer the discussion.

In general, the documents contained relevant information about time constraints, yet how these times were determined for each task was not stated anywhere. It is probably accurate to say that they were determined based on the MUET speaking test format as well.

Table 4.6c

Observation	*Findings*
Researcher	Many students were affected by the time constraints, especially for task A

Commentary

It was apparent that students were working under time pressure, and examiners kept to the times specified for both preparation and presentation stages strictly. Many students expressed dissatisfaction about their performance, especially in task A, as they left. This is one element that the test setters would need to consider revising if they intend to maintain the direct format for testing speaking.

The literature indicates that in the speaking test, different issues present themselves. Examples of these are time for processing normal speech, spontaneous vs. prepared speech (Foster & Skehan 1996), responding to input in an on-line or real time sense (Norris et al 1998), and issues of times for preparation and production; most importantly can the tasks be fulfilled satisfactorily in the time allowed.

4.3.1 b) Test administration

Physical conditions refer to the environment at which the test was conducted in relation to the following factors: Lighting, noise level, temperature, seating arrangements, and conditions for the disabled.

This aspect of the exam is as important as the question paper itself. The test conditions listed above should be similar for all candidates especially if the exam involves a large number of candidates, scattered at different test centres, as is the case with Mara University of Technology, Malaysia. If candidates are taking the same test at the same time but under varying physical conditions, their performance could be affected if the conditions are not conducive. For example, students in the main campus have the advantage of better test venues with minimal noise or other forms of distractions than those in other campuses who suffer from shortage of rooms, or rooms that are

unfavourable for a test. In general, we should aim for non-distressing or adverse physical conditions so that we bias for best in our tests (Weir 2005).

Table 4.7a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	Average of more than 70% agreement on all conditions except 'conditions for the disabled', only 63% agreed
Questionnaire	Lecturer/Examiner (N= 46)	Average of more than 80% agreement on all conditions except 'conditions for the disabled', only 28% agreed
Questionnaire	Administrator (N = 7)	Conditions for the disabled need to be improved
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Those from branch campuses especially, complained about test venues; some held in lecturer's office, noise and interference; in language labs also less comfortable
Interview	Lecturer/Examiner (N = 10)	Test conditions are dependent on the faculty or campus you teach at; branch campuses determine their own conditions, lecturers look for suitable venue for their students
Interview	Administrator (N = 6)	At branch campuses, the speaking test is not given priority
Interview	Expert (N = 5)	No special venues for speaking test; space is a big problem at the university

Commentary

Student data between the questionnaire (quite positive) and interview (rather negative) are contradictory; however, it may be that when they were interviewed, the students had more to say about the test conditions. Like the lecturers/examiners, students from the main campus confirmed that conditions in some faculties were better than others, for example, some of them were in new buildings where the rooms had new furniture, air-condition that did not make loud noises, and where general noise level was low.

Students in the branch campuses had more complaints about where the tests were conducted; the tests were conducted during term time, mostly in lecturers' rooms and offices, and there were other distractions like noise and uncomfortable seats.

It is probably not surprising that this aspect of the test is an administrative concern, and in this survey, the administrators expressed their frustrations on the sheer logistics of managing a large-scale speaking test, especially with the test venue, at the main campus as well as the branch campuses. This had an impact on all participants, especially the students. Moreover, limited considerations or in some centres, none at all, were made for students with disabilities, whether mental or physical in nature. (see Cambridge UCLES on-going work on conditions for disabled candidates in their exams, e.g. Gutteridge 2003, 2006; Taylor & Gutteridge 2003)

Hence, physical conditions are not equal for students in this test. This would affect performance and result in unreliable test scores.

Table 4.7b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	N/A
Question paper	N/A
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	Stated but not in specific terms; lecturers are responsible for organizing the test venue

Commentary

The documents did not specify information about test venues because the class lecturers themselves have to organize this for their students. In the absence of any explicit

guidance for the lecturers, this aspect of the test administration was essentially uncontrolled, representing a serious threat to the validity of the test.

Table 4.7c

Observation	*Findings*
Researcher	Most students were informed of the test venues before the test; very few students were not able to find the test venues. Occasionally at the branch campus, students would gather in their classrooms and given further instructions about the test venue and roster.

Commentary

It was observed that with some exceptions, most test sessions went on smoothly. However, it is the university's responsibility (see table 4.11a above for administrators' testimony) to ensure that testing conditions are appropriate and fair for all students so that many students are not disadvantaged because of poor testing conditions.

In general, test conditions are not suitable, many test participants have expressed frustrations about this, and this can cause variances in student performance, that is, construct irrelevance.

Uniformity of administration refers to a testing environment where the test is conducted according to detailed rules and specifications. This is especially important for large scale or high stakes tests which are conducted across centres and at different times. If the uniformity rule is broken such as if examiners at one centre give extra time to its candidates for planning, or even speaking, then theory-based validity is compromised. It is imperative that this aspect of the test is adhered to strictly, especially in the case of Mara University of Technology, where the test is conducted across many locations throughout the country within a period of one week.

Table 4.8a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	80% agreed examiners followed rules strictly
Questionnaire	Lecturer/Examiner (N= 46)	90% agreed examiners followed rules strictly
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Comments on examiner mannerism, second examiner's presence/absence
Interview	Lecturer/Examiner (N = 10)	- Some cases only one examiner - Students allowed to discuss at preparation time for task B
Interview	Administrator (N = 6)	The rules & regulations specify the conduct of the test; examiners to adhere strictly
Interview	Expert (N = 5)	

Commentary

The agreement on uniformity of test administration is high for both students and staff; however, there were some inconsistencies mentioned during the interviews with both parties. Students were unhappy with some examiners whose mannerisms were intimidating, being their class lecturers, and others who were not helpful (see Brown & Lumley (1997) for a useful account of factors that make a speaking test easy or difficult). Some lecturers were aware of the fact that many test sessions, especially in the branch campuses, had only one examiner present; administrators at those campuses maintained that the problem is one of timetabling, and they entrust lecturers to organize the times themselves. Another inconsistency is some examiners allowing students to discuss amongst themselves during preparation time for task B; this is unfair to students in test centres where it is not allowed. Administrators indicated that their guidelines for

conducting the speaking test were clear and it is up to the coordinators at the respective centres to see that they are followed strictly.

Table 4.8b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Test specs indicate times for preparation and presentation of tasks; content; organization, etc.
Question paper	Indicate times for and order of preparation and presentation for both tasks
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	Instructions for how test is conducted

Commentary

Most documents contained information necessary for examiners to be able to conduct the test in an organized manner. The Test specifications described the test context and other organizational features, although the information is few and brief. The Guidelines and instructions for administering the test has a list of points such as adhering to time constraints for preparation and presentation of both tasks and examiners providing assistance only where they think appropriate, and so on.

Table 4.8c

Observation	*Findings*
Researcher	Most sessions which had only one examiner present were at the branch campuses; examiner mannerisms were overall consistent with very few who provided extra help or none at all

Commentary

Inconsistencies in test administration influence students' performance because it can affect their temperament and cognitive processing during the test. From the observations, it was clear that test anxiety and tensions were high amongst candidates, in spite of the fact that they were familiar with the test environment and know the other speakers as well as at least one examiner.

In general, although lecturers, examiners and administrators were aware of inconsistencies across faculties in the main campus and at the branch campuses, this aspect of the test is not addressed efficiently. In essence, because the test has been carried out for six years without too many major problems, issues like uniformity of administration are confined to the classroom level where, like the physical conditions (above), lecturers are responsible for their students' welfare during the test period.

Security involves limiting access to the specific content of the test to those who need to know it for test development, test scoring and test evaluation.

Test security is important especially for secure tests; test items of these tests are not published and unauthorized copying is illegal. If tests are not secure, candidates may know the answers in advance, and this will affect their processing, i.e. they will rely only on memory. In the case of the UiTM test, security is crucial since it is large scale, spread across more than ten campuses and is conducted within a limited time to thousands of candidates.

Table 4.9a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	48% agreed that students are not able to discuss test questions
Questionnaire	Lecturer/Examiner (N= 46)	36% agreed that students are not able to discuss test questions
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Lecturer/Examiner (N = 10)	Students have a chance to discuss questions with friends, even between campuses; security is a problem
Interview	Administrator (N = 6)	6 parallel sets of question papers; questions for a particular day stipulated by Language centre
Interview	Expert (N = 5)	Difficulty in ensuring when examiners are involved in part of administration & rating

Commentary

Many students and staff disagree that students were not able to discuss test questions because they do; there is no standard procedure to ensure test security, especially in between test sessions within a day or the week. This is possible because many groups are tested in a day and only two sets of papers are used. Lecturers confirmed that they could not prevent this from happening, even though, in addition to following regulations stipulated by the Language centre, examiners required students to dispose of their written notes before they leave the test centre. While administrators ensure that they conduct random checks occasionally (this is done only in the main campus), and that regulations are clear, security is a problem for this test. Administrators at branch campuses complain that occasionally, the question papers get to their campuses late; as a result, they take longer than the stipulated time to conduct and complete the test.

In general, security for a large scale test like the speaking test in UiTM is a problem which the test administrators have not been able to address effectively.

Table 4.9b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	N/A
Question paper	6 sets of parallel question papers to be used within a week of testing
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	Instructions stipulate the questions to be used each day throughout the week and how the test is to be conducted; examiners to follow strictly

Commentary

Each time the speaking test is administered (twice a year), 6 sets of parallel papers are prepared, i.e. six questions/situations for speaking, and these are used throughout the week; they include a range of topics from student activities/experience on campus to more current issues. Evidence of question papers and instructions/guidelines for administering the test were found for each year; content of test papers varied each year and instructions were specific and rather detailed, but have been the same for the past three years.

Table 4.9c

Observation	***Findings***
Researcher	Examiners used questions as stipulated per day; occasionally examiners decide on the topics to use; after a test, students discuss the test with friends waiting for their turns

Commentary

Security is a major problem for this test in spite of parallel question papers and strict regulations. This is one aspect where both students and lecturers agree because they are aware of what really transpires during the test sessions; lecturers have to proceed with conducting the test throughout the day and week with very little time to ensure test security, and students have the advantage of talking to each other before and after the exams. In general, security is a problem surrounding this test.

4.3.2 OVERALL COMMENT ON TASK SETTINGS

We can conclude that for task setting, the problem areas of this test include weighting for each task, known criteria for rating, time constraint especially for task A, uniformity of test administration, physical conditions at some test centres, and test security.

One of the main concerns here is that these problems arise each year. Data reported above indicate several reasons, the main one being that the test was developed parallel to the national MUET exam, especially for response format, order of items, known criteria and time constraints. Test administration for instance, is an administrative concern but in the UiTM test, this responsibility was passed on to lecturers who have to organize the schedule, venue and the assignment of a second examiner. Unlike the national test which has its own contextual and rating descriptors for the general population, (though this information is scarce & not readily available), the UiTM test should have its own documentation with clear descriptions of test tasks, administration and rating, appropriate for its own needs.

More importantly, it should be clear to all parties involved that these aspects of test were developed based on a theoretical model or framework for assessing spoken language, or sound research based on theory and expert judgment, not just based on the MUET with minor adjustments made to suit the candidates. This information was not evident anywhere in the survey, and lack of understanding or confusion amongst members of staff on these aspects of the test is cause for serious concern. Its strengths in test purpose, response format and the order of items were mainly due to participants' knowledge of them from class practices and exposure rather than a clear understanding of what is expected of them in terms of these elements in the test.

4.3.1 c) Task Demands

The demands of the task include both in terms of linguistic input and output from the test, i.e. discourse mode, channel of communication, nature of information in test text, topic, and length of test, lexical, structural and functional discourse, and interlocutor variables such as gender, speech rate, accent, number and acquaintanceship. These are test input which are explicit in their demands and challenge the candidate's ability to comprehend and process the information in order to produce speech, both long turn and short turn in nature.

Linguistic (Input & Output) refers to linguistic parameters found in the test text as well as those expected of the candidates in their performance

Discourse mode in speaking refers to whether the task involves a monologue or interaction in which more than one speaker is involved, such as in a formal interview or

less formal interaction between peers. The important consideration here is the functional range elicited which may vary distinctly between them, such as reciprocity in interaction which is dependent on the relation between the speaker and listener.

In the UiTM speaking test, the discourse mode involves both an individual presentation (monologue) and a group discussion (interaction). Since the speech functions elicited in long turns are different from functions expected of short turn interactions (see O'Sullivan, Saville & Weir 2002, Luoma 2004 on communicative functions of speech), candidates would be expected to produce them accordingly in the two tasks. Hence, in task A, candidates are expected to present factual information through describing, elaborating and/or providing examples; in task B, argue on several ideas by agreeing/disagreeing, persuading and so on in order to conclude with a decision.

Table 4.10a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	64% students agreed, i.e. monologue & discussion, & argumentative in nature
Questionnaire	Lecturer/Examiner (N= 46)	Only 45% agreed on this mode
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Not very certain about this but find task A more demanding
Interview	Lecturer/Examiner (N = 10)	Find both tasks demanding because students are required to demonstrate high level processing within a short time, in spite of link between tasks
Interview	Administrator (N = 6)	It's logical because task B incorporates information from task A
Interview	Expert (N = 5)	

Commentary

Staff response to the item on this element of the task demand is less positive than the students'. This is reflected in the interviews in which they expressed concern that the tasks were too demanding for candidates, especially given the time constraints which are inappropriate for candidates to be able to fulfil the tasks well; the tasks require candidates to demonstrate critical and analytical thinking, not just language ability & content knowledge. Students, however, felt that while task A was more difficult to demonstrate than task B initially, it did not affect them very much because both tasks are related in terms of content, but not necessarily in language. They were aware of the modes for discourse but less so of the related language and functional demands of the tasks.

Obviously, lecturers/examiners felt that in general, candidates were not able to perform as well; processing was affected both internally and in terms of the external resources that they had to draw upon. Specifically, it could be said that lecturers/examiners themselves were not clear of whether the discourse mode was suitable for the test.

In general, there is uncertainty about discourse modes for the test, i.e. whether it is logical to have task A then B because of the link between them, or whether they are appropriate for the purpose of the test; this is evidence of lack of information relating to the test design and the underlying basis for the tasks.

Table 4.10b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	Described briefly here
Question paper	Evident here
Score sheet, scoring guide/rating criteria	Evident in terms of how each task is marked
Instructions/guidelines	How each task should be administered and marked

Commentary

It is clear that in all the documents the test tasks are described according to the objectives of the test. In general, this is beneficial for the lecturers/examiners as they are preparing the students for the test, and students benefit from the clarity of the task descriptions in the syllabus and question papers. However, like the data in Table 4.10a above, the information here is not explicit in terms of sound basis for why the discourse modes were selected for the test, which in turn causes a lack of or varied responses from the participants regarding this element of validity.

Table 4.10c

Observation	*Findings*
Researcher	Students appear comfortable with the modes of presentations, in spite of anxiety & nervousness; no extra help was needed from examiners

Commentary

In spite of test anxiety and nervousness about the speaking test, candidates seemed at ease about the tasks during the tests. This was confirmed by information from the interviews in which many indicated that they had been exposed to the different discourse modes during class practices and from reading materials. However, other factors such as time (Table 4.6a above) and interlocutor variables (Table 4.16 below) played a big part in the outcome of their performance during the test.

Channel of communication has to be appropriate for the purpose of the task as it can have an obvious impact on the performance in the speaking test; candidates who have

to simulate a telephone conversation with an interlocutor in a different room face different pressures from those who speak to an interlocutor face to face in the same room. For the UiTM test, the channel of communication is face to face with all candidates being present in the room during both tasks. However, in the questionnaire survey, this element is also expressed in terms of how test instructions/the tasks are conveyed to the candidate, i.e. in written form (question paper) as well as orally (examiner).

Table 4.11a

<table>
<tr><td>Participant Data</td><td></td><td></td></tr>
<tr><td>Instrument</td><td>Informant(s)</td><td>Observation(s)</td></tr>
<tr><td>Questionnaire</td><td>Student (N= 387)</td><td>More than 80% found written instructions helpful & information was sufficient</td></tr>
<tr><td>Questionnaire</td><td>Lecturer/Examiner (N= 46)</td><td rowspan="3">More than 70% found written instructions helpful & information was sufficient, but many preferred instructions on video, or audio delivery while candidates read the text</td></tr>
<tr><td>Questionnaire</td><td>Administrator (N = 7)</td></tr>
<tr><td>Questionnaire</td><td>Expert (N = 9)</td></tr>
<tr><td>Interview</td><td>Student (N = 64)</td><td>Most were comfortable with face-to-face method; also too many words/long instructions in a page of the question paper</td></tr>
<tr><td>Interview</td><td>Lecturer/Examiner (N = 10)</td><td rowspan="3">Too many words in a page of the question paper; instructions too long, needs to be revised in future tests</td></tr>
<tr><td>Interview</td><td>Administrator (N = 6)</td></tr>
<tr><td>Interview</td><td>Expert (N = 5)</td></tr>
</table>

Commentary

Information from all participants in the questionnaire seems to contradict what they said in the interviews. While a high percentage of students and lecturers/examiners agreed on how the instructions for the tasks were presented, they also commented that

too much information affected how students prepared for the tasks, which in turn affected their performance. Interestingly, students also suggested that instructions for tasks A and B should be separated and not presented on one page, as it is presented now.

In general, students were at ease about the direct method of testing, mainly because they knew the other speakers as well as one of the examiners; they were, however, discontented by the amount of information, presented in written form in the question paper, to process within a short time.

Table 4.11b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Test specs describes mode & channel of communication
Instructions/guidelines	Some mention of how information is related to candidates

Commentary

Since channel of communication, whether in terms of the instructions for the tasks (input), or students' response (output), affects candidates in terms of cognitive processing during the preparation and presentation, it is important that this information is made clear to all parties concerned. More importantly, the way in which the information is delivered to the students, as well as the way in which students are expected to respond to test instructions and tasks, needs to be expressed in clear terms in the test specifications.

It appears that this important information was not available for them.

Table 4.11c

Observation	*Findings*
Researcher	Candidates seemed calm about the test set up, but for some, performance was clearly affected by the presence of the other speakers and examiners

Commentary

In general, students were not poorly affected by the fact that they had to communicate directly to the other speakers; those who were affected, stated later that it was due to language and content difficulty, partly resulting from test instructions that were too long.

Length of the text or test task needs to be appropriate for the target situation requirements of the students. It affects cognitive processing in terms of strategies and skills as well as executive resources needed to fulfil the task.

For the UiTM test, participants made reference to the time it took to complete the test, i.e. two minutes to present task A and ten minutes to complete task B. The length of the text, or test instructions for this group of candidates, was referred to in the section above on 'channel of communication'.

Table 4.12a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	50% agreed task A was long enough; 77% agreed for task B; instructions were too long
Questionnaire	Lecturer/Examiner (N= 46)	36% agreed task A was long enough; 77% agreed for task B; instructions were too much
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Many agreed that test length is important as it affects how much they had to do during the test
Interview	Lecturer/Examiner (N = 10)	Test length is an important consideration for everyone involved in the test, especially students and examiners
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	

Commentary

As stated earlier, participants responded to this element with reference to the time it takes to complete the test. In general, students and lecturers/examiners felt that two minutes is too short for candidates to demonstrate their ability in task A, but were satisfied with the time it takes for candidates to complete task B. Both parties were also dissatisfied with the instructions for the tasks which were long and compact with information; this was mentioned several times in the interviews relating to elements of time constraint, channel of communication and test length (all in tables 4.6a, 4.11a, 4.12a above).

In fact, just as the validity components in the framework interact, elements within the components interact and overlap as well. The length of test tasks affected processing a range of information within a short time and hence, performance was affected too.

Table 4.12b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Test specs describes test length and length of text for test instructions
Question paper	Evidence of length of text/instructions for tasks
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	N/A

Commentary

Only the test specifications and the question paper itself had evidence of how long the test instructions were and how much information was in them. Although students may have seen some past question papers in the classroom, there is evidence during the interview that some had not seen any. Hence, seeing the length and amount of information in the test for the first time during the test would unquestionably affect a candidate's ability to process this information when attempting the tasks.

This is a concern of the test knowledge of students; it is not just enough for students to know about the test topic or format (or weighting; response time or criteria etc.), but they also need to know about what is expected of them linguistically in the test (in terms of input and output).

Table 4.12c

Observation	*Findings*
Researcher	Candidates spent the preparation time reading the instructions and writing notes; the time was insufficient for some who could not complete their preparation, this inevitably affected their performance.

Commentary

It appears that the tasks for this speaking test came with a set of instructions which were long and contained a lot of information for candidates to process within a time that was insufficient for them. It was obvious from the observations that candidates were unhappy about this, especially those whose performances were affected by incomplete preparation, demonstrated by incoherent speech and frequent stops during presentations and interactions.

In general, the lengths of the text or instructions for the tasks do not appear to have been appropriate for this speaking test.

Nature of information refers to whether the information in the text is relatively abstract (e.g. love, death, etc) or concrete (e.g. pictures on the wall); it should be relevant and appropriate for the test task. In general, abstract information may be more challenging to process in terms of its linguistic and cognitive demands. For this speaking test, the aim was to present candidates with factual, experience-based type information, rather than abstract ones, such as a topic on art appreciation or the value of life.

Table 4.13a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	68% agreed that test has both types of factual & abstract information
Questionnaire	Lecturer/Examiner (N= 46)	68% agreed that test has both types of information
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Lecturer/Examiner (N = 10)	A mixture but mainly experience-based type information, rather than factual or abstract
Interview	Administrator (N = 6)	Information is factual and has to be manageable for the students
Interview	Expert (N = 5)	Nature of information should be familiar to the students and within their experience

Commentary

Students had no comments during the interviews as they weren't sure of the type of information in the test; this is an indication of the lack of clarity of the information in the test task. Lecturers/examiners agreed with the students but stated in the interviews that information is usually tailored to what students are familiar with in terms of their experiences at university or knowledge of current affairs. If this were the case, then students need to be informed of what to expect in terms of this element in the test. It appears that they knew of the types of topic for the test but not details of the nature of information they could expect. For example, one of the topics (see Sept/Oct 2004 Speaking paper) used was 'National Service', which was current and new at the time; many of the students had seen the topic in the media but information in the test question was unfamiliar and even difficult to process.

Table 4.13b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Described in test specs briefly, e.g. familiar/academic and/or social in nature
Question paper	N/A
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	N/A

Commentary

It is expected that this information should be found in the test specifications; however, the description found was brief and vague, and lacked clarity. If lecturers are not certain about this information, they would not be able to relate it to the students, and students will not be aware of the nature of information in the test until they are faced with the question paper during the test.

Table 4.13c

Observation	*Findings*
Researcher	It wasn't clear if candidates were affected by the nature of information in the test until the discussion in task B where some were not able to contribute

Commentary

If the nature of information in the test is not explained to students in clear terms, they will not know what content/type of information to expect in the test, and this would certainly affect their performance, as demonstrated especially in task B. Here,

information from another speaker may be unpredictable and a speaker is unable to respond appropriately due to lack of knowledge and unfamiliarity to the information. In general, candidates are not aware of the nature of information of the tasks that they will be faced with in the test, and this probably affected their performance. Hence, it is not only a question of concrete or abstract type information; it is beyond this because the nature of information in the test is often unpredictable and complex for these candidates.

Content knowledge required for completing a particular task affects processing in terms of executive resources (background and specific subject knowledge) that candidates have to draw upon when faced with the task. The relationship between content that the task demands and knowledge that candidates bring to the test needs to be considered; content should be suitable (genre, not bias to any group, etc) & familiar.

Table 4.14a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 387)	61% agreed that content/topic was familiar
Questionnaire	Lecturer/Examiner (N= 46)	55% agreed that content/topic was familiar
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Familiar for those exposed to similar topics in the classroom and had time for extra reading
Interview	Lecturer/Examiner (N = 10)	Familiar because students had researched & discussed similar topics; many still don' tread enough
Interview	Administrator (N = 6)	Main problem is students do not read enough, especially on current affairs
Interview	Expert (N = 5)	The problem concerns all students; not enough effort put into reading on current affairs

Commentary

The low percentages of agreement on this element indicate that in terms of internal and external knowledge, candidates were not prepared. This was confirmed in the interview data where students indicated that many of them did not have enough time to research the topics proposed by class lecturers for practice, and they themselves did not read enough to prepare for the test. Lecturer/examiners agreed with this; students generally do not read enough so they are at a disadvantage in terms of the external resources/ content knowledge they bring to the test. Administrators and experts concluded that lack of exposure to these topics is a major problem among all students, and this contributes to their inability to participate in the discussion during the test.

However, lecturers are required to provide sufficient background in terms of the test content to their students; student interview data revealed that some lecturers do and some do not because of time factor.

It appears that in this aspect, some candidates were disadvantaged because of lack of information and exposure in the classroom. It appears too that there are difficulties in relying on teachers to take the responsibility of informing the students about the content knowledge that is expected of them in the test; this is similar to the problems faced in other elements above (tables 4.3, 4.4, 4.11, 4.13).

Table 4.14b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Brief description in test specs on text type
Question paper	N/A
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	N/A

Commentary

A very brief description of the type of content students should expect is found in the test specifications; it is brief and vague...'Social and academic contexts...and topics related to family, college and community issues'.

In this aspect, there seems to be insufficient information in the documents, which may cause a discrepancy in the amount and type of information that lecturers disseminate to their students about test content.

Table 4.14c

Observation	*Findings*
Researcher	In general, students' performances were somewhat affected in terms of content and delivery.

Commentary

In general, the candidates managed to attend to the tasks fairly well, but many also struggled from lack of points/ideas and seemed affected by the topics and content of the test. This aspect of the test is vital for candidates to be able to process information effectively; some were obviously not able to achieve this. Further interviews with students indicated that they found several aspects of the test that affected their

preparation and performance; some candidates had more difficult topics than others, within a topic for discussion a candidate might be asked to argue a more difficult position/idea compared to others in the group. Hence, this was an unfairness of the test which disadvantaged some candidates more than others. This is linked to the problems of establishing equivalent or parallel forms (O'Sullivan, Weir & Horai 2004) and task difficulty (see Fulcher & Reiter 2003) for a speaking test.

Linguistic variables of text (in main text, test instructions, or task) need to be considered in both task input and task output. In the findings below, the variables are reported together because questionnaire data showed participants' overall agreement on them, but more information was gathered from the interviews. These include:

Lexical range refers to lexical items or words used in the text which are appropriate for the students' ability level. Generally, texts with low-frequency words are more difficult to understand than those with high-frequency words (Weir 2005).

Structural range of the text provided should contain familiar syntactic patterns and less complex grammar. This will enable candidates to process more efficiently while attempting to understand the instructions given to them.

Functional range refers to the illocutionary force of what is said, such as to persuade, describe, and so on. In the speaking test, the instructions/test tasks need to be explicit about what functions the candidates are expected to demonstrate.

Table 4.15a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	More than 70% agreed; found no difficulty in terms of words, structure and functions; they were clear
Questionnaire	Lecturer/Examiner (N= 46)	More than 70% agreed that students had no difficulty with words, sentences structure and functions found in test questions
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Understand most words in the text; can ask for help if not certain of a word, but problem is some had more difficult topics and tasks than others
Interview	Lecturer/Examiner (N = 10)	Students are allowed to ask the meaning of words/phrases; they complained about difficulty in topic range and task within a group, probably related to structural difficulty and lack of content knowledge
Interview	Administrator (N = 6)	Examiners can help without giving away too much of the question
Interview	Expert (N = 5)	

Commentary

In general, most participants agreed that candidates did not have serious problems with lexical and structural range. Lecturers and administrators confirmed in the interviews that when setting the test questions, they avoided low frequency words, phrases/ expressions and complex sentence structures; yet, many candidates had problems understanding the tasks. It could be the functional range of the questions which became unclear as a result of long and complex instructions and descriptions of test situations. Examiners stated that candidates had difficulty in thinking of their points in an organized manner. Candidates complained of difficulty completing several tasks in a short time (2 minutes preparation time); they had to read, analyze and organize their ideas in a coherent manner before the presentation.

In general, although the lexical and structural range of the test was appropriate for the candidates, it appeared that functional range may have caused difficulty for them. This is a result of a set of instructions and task descriptions that were long and loaded with information which resulted in difficulty to determine the functional range of the task. This in turn affects candidates' internal processing in terms of conceptualization of the linguistic demands of the tasks (input and output), the message is misinterpreted or lost, and overt speech is affected.

Table 4.15b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	Brief description in test specs on linguistic range of questions
Question paper	Linguistic range evident in instructions and task descriptions
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	N/A

Commentary

The test specifications for the test give a brief description of the linguistic range that test questions entail. Candidates are also exposed to past test papers to give them an idea of the linguistic features of the test questions. Several of these test papers were inspected & they showed similarities in terms of the linguistic range; if they had access to these papers, candidates would have a better idea of the difficulty level of the questions in this respect. While the linguistic range appears appropriate to the candidates at this level (Intermediate - Advance), the topics and the speech functions are challenging and require more time than was stipulated in the test to ensure that all candidates have an equal chance to demonstrate their abilities. In fact, topics change each year quite

drastically, (e.g. community related activities such as for the RSPCA, crime prevention amongst youths (see Sept 2004 Speaking test), to more serious activities such as effective measures for future disasters like the tsunami, to improve the quality of life for the disabled (see Sept 2005 Speaking test)). While speech functions are maintained each time, candidates seemed to have difficulty with this aspect of the test.

Table 4.15c

Observation	*Findings*
Researcher	In general, candidates were able to cope with the test fairly well, in spite of their grievances about the test questions during the interviews; still, many had difficulty fulfilling the tasks. Candidates were allowed to ask for clarifications on a word/phrase/expression before the test starts; examiners were generally helpful.

Commentary

In general, it appeared that the test sessions went on smoothly in spite of several groups that had serious lexical, structural and functional problems when attempting the tasks, and in spite of data gathered from questionnaires in which they indicated that they had no serious difficulty with the words, structure and functional range of the test tasks. The Interview data, however, showed that many candidates were troubled by the amount of information they had to process and the difficulty in terms of differences between topics and within a topic for discussion. It may, in fact, be the structural and functional range that caused difficulty in comprehension and ability to organize ideas coherently. This is linked to evidence in the document analysis above (table 15b) which showed vague and insufficient descriptors of linguistic variables that the test demands of the students.

Interlocutor variables refer to input dimensions, concerned with features of the language used by the interlocutor and how these affect or influence the performance of the candidate. Weir (2005) stresses that this is the least definable aspect of the framework and may be problematic to put into operation unless an interlocutor frame is employed (see Cambridge ESOL in Weir & Milanovic (eds.) 2003).

In the UiTM speaking test, candidates receive minimal input from the examiners and most of the input from other speakers in the group discussion task. This in itself may be a problem for tests such as this, which is narrow and specific in scope yet broad and open-ended in how the discussion is controlled and managed; the outcome is often unpredictable. The variables are: *Speech rate; Acquaintanceship; Variety of accent; Number* and *Gender*

Table 4.16a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	Except for 'number' (54% disagreed with presence of second examiner), agreement was high (75%-85%, mean 4.0) for speech rate, accent, gender and acquaintanceship
Questionnaire	Lecturer/Examiner (N= 46)	Results different from student data: Accent (59% agreed students had no problem here) Gender (50% agreed this affected students) Number (90% thought students had no problem here). Speech rate + acquaintanceship were similar to student data (> 80% agreement on both variables)
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Acquaintanceship: Important to know who you are talking to, especially since one's discourse in group task is dependent on what others say, and this affects the outcome of the discussion 2nd examiner's presence increased anxiety & affected performance; effect can be negative or positive
Interview	Lecturer/Examiner (N = 10)	Knowing who they are talking to is important for the students Presence of 2nd examiner affected many students
Interview	Administrator (N = 6)	Presence of a 2nd examiner is compulsory to prevent bias and ensure marker reliability
Interview	Expert (N = 5)	

Commentary

Students were not affected by the interlocutor variables, in fact, they were satisfied with the test set up, i.e. one of the examiners is a class lecturer and the other speakers are their classmates. In the interview, students stressed the importance of knowing the people they are talking to because this makes communication easier, causes less stress on the speakers, and is comforting overall. Many were however, unhappy with the presence of the second examiner, who is usually another lecturer they may or may not know. Lecturers/examiners agreed that acquaintanceship is very important for the students; they prefer talking to their classmates than students from other classes, they

seemed more comfortable and more able to control the discussion to their advantage. While they were uncomfortable with the presence of a second examiner, this is a requirement of the test and many students may not understand this if they had not been informed of it. Students however, were not in favour of the 2nd examiner's presence mainly because of examiner's conduct during the test; some made comments during presentations, others made expressions through facial and other non-verbal cues and gestures.

In general, candidates felt they were not negatively affected by the interlocutor variables, especially since they were comfortable presenting to and interacting with people they know. Except for acquaintanceship, lecturers/examiners had differing viewpoints; they felt students were affected by gender and accent, but not number.

Table 4.16b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	Information on interlocutor variables not available in any of the documents
Question paper	
Score sheet, scoring guide/rating criteria	
Instructions/guidelines	

Commentary

There was no mention or description of interlocutor variables in any of the documents. This could be either because lecturers are expected to explain them to the students or they were not considered important enough to be listed in the test specifications or guidelines. In fact, these are variables that could affect student performance (as the data above shows) and they should be listed or described clearly in the test specifications

and/or guidelines. Students need to be given sufficient and accurate information regarding interlocutor variables, what they are and how students are expected to perform under these conditions. Interlocutor variables can affect both planning for the task and monitoring their speech during the presentations. (see Levelt 1989, chapter 2 for detailed account of the role of interlocutor in conversation)

Table 4.16c

Observation	*Findings*
Researcher	In all the test sessions, input from examiner(s) was minimal, only brief oral instructions to students before the test; they are not allowed to provide help at other times. Students appeared comfortable talking to and with their classmates, though the amount of actual interaction was minimal. In the branch campuses mainly, only one examiner was present instead of two, usually this would be the class lecturer.

Commentary

In general, students were apparently not affected by these variables in a negative way, and seemed capable of coping with speech rate, accent, number and gender of other speakers during the test. This is not surprising because the candidates are relatively homogeneous in terms of race, native language, level of ability and English is a second language for most of them. As the data above (Table 4.20a) shows, students agreed that acquaintanceship is a major factor for them in this test; they are comfortable talking to people they know, including examiners, and this affected their performance in a positive way. (see Weir 1993 on interlocutor status & familiarity; O'Sullivan 2000a, 2000b, 2002 on acquaintanceship; Berry 1993, 1994, 1996 on interlocutor personality)

4.3.3 OVERALL COMMENT ON TASK DEMANDS

We can conclude that for task demands, the problem areas of this test include length, nature of information, content knowledge required and functional range. There seemed to be fewer problems concerning discourse mode, channel of communication, word and structural range, and interlocutor variables, especially acquaintanceship.

However, data above also indicated several points in which confusion and overlap of information occurred, such as channel of communication (table 4.11) and test length (table 4.12) which were frequently tied to time constraints. Like the conclusions for test setting above, most elements for task demands were not described and explained in explicit terms in the test documents to the candidates, especially nature of information, content knowledge required, and lexical, structural, functional range of the test tasks, and interlocutor variables. Like rating criteria and order of items (for test setting), the lecturers were inconsistent in providing students with information and details relating to content knowledge and nature of information. More importantly, students also need to know about what is expected of them linguistically in the test (in terms of input and output), how these variables can have an effect on how they prepare for the tasks and on their performance; there was no evidence of this in the documents and from interviews.

Furthermore, it was also not evident in the documentation and from interviews with administrators and experts, how these variables were determined; whether research had been conducted, expert judgment and language theories consulted were not clear. In

any test, especially for a speaking test of this nature and scope, many considerations need to be well thought-out, not least the test task, test specifications and rating scales. (see Fulcher 2003, chapters 2-5 on developing the speaking test).

All these factors: insufficient and accurate information in the documentation, lecturers' inability to provide relevant test information to students, and how these elements of the test were determined, contribute to lack of validity in terms of task demands for this test.

4.3.4 CONCLUSIONS FROM CONTEXT VALIDITY ANALYSIS

In conclusion, it can be argued that the test lacked context validity in many aspects of the validity component and this is related to test knowledge on the part of members of staff involved in developing the test, content of test documentations, and theoretical basis/model from which these elements in the test were determined.

It was evident especially in the interview data (see Appendix 3.5: Staff interview) that many staff participants have knowledge of the test from experience teaching the course and conducting the test. Very few indicated that they know and understand the elements of the speaking test and factors that affect students' performance, not from test specifications or the guidelines/instructions to the test, but mostly from experience. While most lecturers were able to conduct the test as instructed, they lacked valuable knowledge needed to enable candidates to maximize their preparations for the test and to provide all candidates an equal chance of demonstrating their speaking abilities

during the test. For example, interviews with staff indicated many were not clear on why it is important for rating criteria, weighting of tasks and time constraint to be indicated explicitly in the question paper. Inconsistencies were also found in terms of lecturers' reliability in disseminating information to students. Students received either unequal amounts (see tables 4.3, 4.4, 4.7, 4.8, 4.9) or quality of information (see tables 4.10 - 4.115) regarding these test elements. Students need to be informed in clear terms what the elements are and how performance can be affected by these conditions, i.e. how they are expected to respond to the setting and linguistic demands (input and output) of the test.

Related to this is that documentation of these important elements are either brief or vague and not explained in complete terms. Documentation on purpose, nature of information and text type/length were found in the test specification, which is just one page long and is the only document describing test content. (see Appendix 3.7b). As mentioned above, students not only need to know about these parameters, but more importantly how they are expressed in the test and how students are expected to respond to them accordingly. Test writers, administrators and lecturers need to have sufficient background knowledge on what these elements mean to the students if the test is to affect test performance in a positive way.

Finally, it is also evident from both documentations and interviews with administrators and experts that the test was not theoretically driven or developed on a clear model/framework for testing spoken language (see Weir 1993, Bachman & Palmer 1996, Tarone 1998, Fulcher & Reiter 2003, Fulcher 2003, on important considerations for

designing speaking tests). Without a clear model in which the test was designed and developed, staff members lack adequate understanding of the test and candidates suffer from this deficiency. This results in a test that is deficient in its context validity, and this inevitably affects test taker's internal processing in terms of cognitive complexity and familiarity, code complexity and communicative stress (from Skehan 1996). Scoring validity is also affected in many ways, especially since the criteria for rating is derived from linguistic demands of the test (see data Scoring validity below),

4.3.5 THEORY-BASED VALIDITY

The aim in this part of validation is to understand test takers' cognitive process of and affective response to the direct method of testing spoken language. This aspect of validity has to do with the 'how', or the operations involved when test takers attempt the task, hence the internal processing. Bachman's (1990) communicative language ability model included both language competence and strategic competence, equivalent to the executive resource and executive process that test takers have to access during the test that appear in Weir's (2005) framework (see Figure 4.5 below).

It is also important to consider the characteristics that test takers posses at the time they undertake the test. The test taker's characteristics (O'Sullivan 2000) are directly related to theory-based validity as these characteristics will directly impact how the test taker processes the test task, i.e. test takers bring inherent characteristics to the test and these influence the internal processing that they undergo in order to attempt the task. (for an overview of the characteristics see Weir 2005, Chapter 5, pp 51-55).

Figure 4.5 Aspects of Theory-based Validity for Speaking (from Weir, 2004)

THEORY-BASED VALIDITY		
INTERNAL PROCESSES • Conceptualiser • Pre verbal message • Linguistic formulator • Phonetic plan • Articulator • Overt speech • Audition • Speech comprehension	M O N I T O R I N G	EXECUTIVE RESOURCES **Content knowledge** • Internal • External **Language knowledge** • Grammatical • Discoursal • Functional • Sociolinguistic

Internal processing is divided into executive process and executive resources.

For speaking, Levelt detailed this process in his blueprint for speaking (Levelt 1993, chapter 2: 6-9). The processing system encodes a message from the time it is conceived until it is produced as preverbal message (a conceptual structure), which is then translated by the formulator into a linguistic structure through a series of grammatical and phonological encoding. This 'internal' speech is articulated as overt speech by the musculature of the respiratory, the laryngeal, and the supralaryngeal systems. A speaker is also able to listen to and monitor his own overt speech just as he can listen to the speech of his interlocutors.

Data for this aspect of validity is reported below in terms of the above description in two parts:

Executive Process involves: a Conceptualiser phase; Linguistic formulator in terms of words/expressions and structures of the test text; Overt speech in terms of strategies for appropriacy, accuracy and organization of the presentation.

Executive Resource includes: Content knowledge (internal & external) and Language knowledge which consists of grammatical, discoursal, functional and sociolinguistic in nature.

This is the section of the framework that relates to assumptions that test developers, writers, experts etc. make about students' ability and knowledge of spoken language, what processes they should employ and knowledge they should have (internal, external, linguistic) to be able to perform the tasks well.

In terms of this study, it was difficult for candidates to interpret and express what they did in the test to fulfil the task. For these candidates, this was their first time participating in a study of this nature. However, the data below were gathered from candidates through questionnaires and interviews after they had completed the direct speaking test. In addition, as reported in chapter 3 (see section II for the development of the questionnaire and the interview framework), the research instruments were developed over the span of many months, which included trials and pilot studies. Hence, in addition to the extensive preparation involved in developing the research instruments, this was the first time a study of this nature, i.e. gathering validity evidence on an existing test based on a framework was carried out at the university in Malaysia and for all those involved in the process.

The following are data from two parts of internal processing which students were expected to go through during the speaking tasks.

4.3.5 a) Executive process involves the following:

Conceptualizer/Preverbal message

Table 4.17a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	90% agreed they read carefully, thought of points and wrote down points; 70% agreed they thought of satisfying examiner
Questionnaire	Lecturer/Examiner (N= 46)	78% agreed they read carefully, 90% agreed they thought of points and wrote down points; 35% - 50% agreed students 'thought of satisfying examiner' for both tasks
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	For preparation: from mental preparation, prepared notes, mapping ideas, translate L1-L2
Interview	Lecturer/Examiner (N = 10)	Insufficient time to prepare well: range from writing notes to mental preparation, many translating L2-L1 and then think in L2
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	

Commentary

Data showed that students demonstrated some form of conceptualization in terms of how they prepared for the task. There may be other processes not listed in the questionnaire and this was evident in the interviews when some students mentioned strategies such as mental preparation, mapping down their ideas, and many who had to internalize the instructions and task in their L1 before they could translate them and present in the target language. They did not 'think about how to satisfy the examiner' possibly because of the time pressure they were under; in fact, some stated in the interview that they did not think of anything but to get straight to the task at hand. Although lecturers/examiners agreed with the student data on two elements (thought of and wrote down points), they disagreed strongly that students had read carefully and thought about how to satisfy the examiner. They also believed that many students were

not able to prepare for the task effectively because they would translate the instructions and task from English to their mother tongue and write down a lot of notes; very few were able to prepare mentally or organize their thoughts well. However, students showed better preparation for task B because they had listened to the points presented in task A.

In general, it is clear that students had thought of how to attend to the tasks at the preparation stage by employing some degree or level of internal processing of the input, i.e. at the conceptualization phase. Data for task B showed slight improvements in numbers and responses because for the second task, the conceptualization stage involved incorporating information from task A, and this may have made planning more manageable.

Table 4.17b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	No mention or description of internal processing required for task fulfilment in any of the documents
Question paper	
Score sheet, scoring guide/ rating criteria	
Instructions/guidelines	

Commentary

It is probably common not to find this information in the specifications for many tests, but some information regarding this aspect of theory-based validity would be a big advantage to candidates. Lecturers could provide students with detailed information on what the tasks entail in terms of what they need to think about or processes they could

employ to achieve the task. They, of course, would need to have knowledge regarding the processes, for example as detailed by Levelt, in order to impart the information to the students. If they do not have this knowledge, a description or some notes on cognitive processing in the test specifications or guidelines would be helpful for them and the students.

Table 4.17c

Observation	*Findings*
Researcher	Most students were able to prepare on time, and this was reflected in their performance; those who were not able to, showed problems in their presentations and contributed minimally to the discussion

Commentary

In general, many students were able to prepare for the tasks on time, though there were many who complained of the time constraint (also evident in table 4.8 – 4.10) in context validity above). It is possible that due to time pressure and the amount of information, internal processing at the beginning of this test, i.e. preparation stage, was affected and this in turn affected performance. They had to conceive the information, apply analytical and organizational skills for the content, and translate these in terms of the grammatical and phonological encoding, in preparation for the actual presentation. Though these processes happen at a sub-conscious level for the candidates, knowledge on key processes of conceptualization would facilitate them at the planning stage.

Linguistic formulator

Table 4.18a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	70% - 80% agreed they thought of words/ expressions, and structures needed for the task
Questionnaire	Lecturer/Examiner (N= 46)	> 90% agreed students thought of structures needed, more in task B than task A (41%); the reverse for 'thought of words/expressions', i.e. higher for task A (64%) than task B (50%)
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Did think of words and structures to use; also listening to presentations in A important for preparation in task B
Interview	Lecturer/Examiner (N = 10)	Due to time & anxiety, students not able to think about language; writing notes and translating ideas a lot
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	

Commentary

Data from student and staff were contradictory to each other; students showed high agreement on whether they thought about words and structures to use for both tasks, but lecturers/examiners disagreed strongly that they did this, especially in task A. In fact, staff data indicated that the lecturers felt for task A students thought more of the words/expressions than about the structures to be used, and the reverse for task B where they felt that students thought more of structures. This is possibly due to the nature of task B where there is interaction involved, though the connection was unclear in the findings. Students stressed in the interviews that they thought hard about words and structures they had to use, though the outcomes can be different from what they had planned. In addition, drawing on points from A was important in preparing for

task B, hence, students were aware that listening to others present in task A is crucial for task fulfilment in B. Examiners, on the other hand, felt that students were not able to think of language for the tasks due to time pressure and anxiety; they spent time writing notes and translating ideas from L2 to L1 and vice versa.

In general, there seemed to be evidence that students employed the linguistic formulator to determine the extent to which their word use and sentence structures meet the requirements of the tasks, even though the lecturers/examiners maintained otherwise.

Table 4.18b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	Brief note on the language used in input (Formal/semi-formal)
Question paper	Evidence of the language used in instructions + task description
Score sheet, scoring guide/rating criteria	N/A
Instructions/guidelines	N/A

Commentary

As mentioned in the Conceptualizer/preverbal message section above (table 4.17), it would be useful for the test specifications to describe in detail the type of language used in the input, and language necessary for students to fulfil the tasks; this is not evident in any document.

Table 4.18c

Observation	*Findings*
Researcher	Students had problems with language use, especially in task A; they had less language difficulties in task B, other speakers helped/spoke when someone is stuck

Commentary

It was clear that there were more students experiencing language difficulties in task A than in task B; in task B students seemed more relaxed possibly because there were other speakers who could 'help out' or speak when a student was having difficulties on a point or on language use. However, all this evidence shows students undergoing some degree of internal processing in an attempt to form the language necessary for their presentations in the test, at the lexical as well as the structural level.

Overt speech

Table 4.19a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	60% - 70% agreement on **checked** word use & organization, 40%- 50% agreement for grammatical accuracy, 72% on effect of what they said; average 70% agreement on **adjusted** word use & organization, 60% on grammatical accuracy; > 80% agreed they checked points to make, other speakers' points & adjusted their response based on what others said
Questionnaire	Lecturer/Examiner (N= 46)	Percentages much lower here: Average 30% agreement that students **checked** word use & organization, below 20% on students checked grammar; even lower agreement on **adjusted** these elements, especially grammatical accuracy; but high agreement that students checked points, adjusted their response, etc.
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	13 instances recorded on some 'monitoring' of their speech, especially for language (word use & grammar); also listening to what others say + how each person interacts can be advantage or not for the discussion
Interview	Lecturer/Examiner (N = 10)	Other factors affect their speech: time pressure, their own language ability, and personality differences, e.g. students from rural areas show less enthusiasm and participation
Interview	Administrator (N = 6); Expert (N = 5)	

Commentary

Student data shows agreement that they monitored their speech and to a certain extent the speech of other speakers, especially in terms of language when they check and adjust their own language use, and listen and respond accordingly to what others say. Lecturers/examiners show much lower percentages in that they felt that students were not able to monitor their language use effectively due to time constraints and other factors. These are found in the interview data (both student & lecturer/examiner), such

as interference from the first language, their own language proficiency, and even the maturity to think through the task in an organized manner so that arguments can be presented coherently and with justifications.

In general, there is evidence to show that students perform some level of monitoring during their presentation and discussion. They are aware of some of the errors they made and corrected them. They are also aware of the points or arguments other speakers make so that they can respond accordingly, or even adjust their responses; they monitor not only for meaning but also for linguistic well formedness (Levelt 1993).

Table 4.19b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	No description of the intended or expected output from candidates in terms of content and language
Question paper	N/A
Score sheet, scoring guide/ rating criteria	N/A
Instructions/guidelines	N/A

Commentary

Again, upon inspection of the documents above, there was no evidence of descriptions of the intended or expected output from candidates in terms of content and language. The only description available is of the functional range of the tasks, e.g. 'Using the right expressions in individual presentation and group discussion such as: making recommendations, stating and justifying opinions, presenting alternative points of views...' (see Appendix 3.7b: Test specification/ Speaking)

Table 4.19c

Observation	*Findings*
Researcher	Students displayed some degree of monitoring, especially in task B, in their own speeches as well as when responding to others. They were conscious of what they said and were sometimes alert to what others are saying

Commentary

Largely, students were well aware of what they had to do and what they were doing during the presentations. In spite of earlier comments about time constraints especially for preparation, most students were able to perform fairly well in the test. They showed some level of monitoring of their own speech, e.g. making corrections of word use or grammatical errors, and the speech of other speakers such as restating a wrong word or use of expression, and even helping to explain or elaborate on a point in the discussion. In general, students were able to demonstrate some degree of awareness on the importance of listening to others (audition) in order to respond appropriately and monitoring their speech in terms of meaning and language use during the presentations.

4.3.6 OVERALL COMMENT ON EXECUTIVE PROCESSES

We can conclude that for executive processing, students showed some level of conceptualization in the preparation stage of the test by use of linguistic formulator to determine the extent to which their word use and sentence structures meet the requirements of the tasks. Students may not be able to express these processes verbally and explicitly, but evidence above (Table 4.18, 4.19) shows that they are conscious of some aspects of them to a certain degree when they corrected their own errors at the word or structural level.

There is also evidence here of the co-construction of discourse, with students indicating that their responses, particularly in task B, are at least in part influenced by what their peers say (table 4.19a) . While literature states the group discussion as being quite popular especially for promoting language use and acquisition, (Long & Porter 1985; Long 1989; Gass & Varonis 1985; Duff 1986; Rulon & McCreary 1986; Berry 1993/1994; Fulcher 1996; Bonk & Ockey 2003), it is also seen as a potential problem with this type of speaking test. Because communication amongst candidates is co-constructed, i.e. depending on each others' response to carry out a discussion, effective communication can only be achieved if this is realized in classroom practices and training. (see Brown 2003; Swain 2001; McNamara 1997; Hughes 2002; Fulcher 2003; Luoma 2004; Dimitrova-Galaczi 2004). In an interview task, interviewers are trained on how best to conduct a fair and balanced interview in terms of agenda management and reciprocity; in the group task, candidates could be taught interactional functions and agenda management expected of them in an interactive task. In the test above, though students show some awareness of the co-construction of their discourse, there is no evidence from all participants and the test documents to indicate that this is a major concern in the group discussion task; because there is no information available, and participants do not have knowledge about it, students are again disadvantaged in task B.

More importantly, for such interaction-based tests, the test design itself should spell out clearly and unambiguously conditions and operations relevant to the construct, and this unfortunately was missing in the test documents (Tables 4.17b, 4.18b, 4.19b).

Finally, it is noted that the lecturers' reactions to students' performance in the test in terms of preparation and presentation were usually contrary to the students' testimonies. They felt that the students did not process the tasks well during preparation, for example, in terms of thinking about the points carefully, word, structures and organize ideas coherently, and during presentation such as checking and/or adjusting their word use and structures. This is interesting as it suggests the limitation with asking lecturers to make predictions of what their students can or cannot do in tests of this kind. Perhaps this testimony should be expressed better by the test developers, experts, etc, but our data from lecturers who were involved in test writing, experts and administrators showed the same results (tables 4.17a, 4.18a, 4.19a above).

4.3.5 b) Executive resources

Content knowledge consists of : Internal knowledge or the test taker's prior knowledge of topical or cultural content (background knowledge) and external knowledge or knowledge provided in the task.

Table 4.20a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	Percentage in the 50s on topic and information familiarity from previous reading and experience 70% agreed that information provided in the test were necessary for them to be able to fulfil the tasks
Questionnaire	Lecturer/Examiner (N= 46)	64% agreed topic was familiar for students; 41% agreed information was familiar from students' previous reading and experience More agreement for information in task A(77%) than task B (50%) that is useful/helpful to students
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	30 instances recorded on topic familiarity: some had exposure in class when topics were discussed, some from media & personal experience, e.g. on qualities of a student leader
Interview	Lecturer/Examiner (N = 10)	Familiarity depends on factors such as where a student comes from, rural or urban, which will determine exposure to the topics and opportunities for reading & experience
Interview	Administrator (N = 6)	Test setters reminded to prepare topics and information that are not biased, familiar and within students' experience
Interview	Expert(N = 5)	Topics relate to current issues and activities that affect student lives in campus and outside

Commentary

Though questionnaire data for both participants recorded low percentages on whether students were familiar with the topics and information in the test, interview data showed otherwise. Many students said the topics and information were familiar either from brainstorming and discussion sessions held in class before the test or from their own experiences of campus life or life as a student in general, or from the media, though this might be true for students in the main campus only. The information provided in the instructions and test task (external knowledge) were necessary and sufficient for

them to be able to fulfill the task. Lecturers felt the topics were familiar to the students for the same reasons they had cited, but many of them obtained fairly low marks because they lacked content and a well-organized presentation; they did not read up and research the topics adequately. Administrators and experts indicated that they tried to ensure that test developers are constantly reminded to include topics and information that students can cope with, not biased to any group, and are within their experience or exposure.

Thus, students may not have sufficient internal knowledge to do well in the test but for them, the test provided adequate amount of information to complete the tasks. However, this does not ensure that they will perform well in the test as the processing theory states the importance of recourse to both internal and external knowledge for discourse to succeed. On the contrary, if information from the test is sufficient for students to fulfill the tasks well without help from their executive resources, the test had not fulfilled theory-based validity. In addition, lecturers feedback showed students who obtained higher scores were those who displayed knowledge of the topics and had understood the tasks well.

Table 4.20b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Brief mention of nature of topics: General topics related to family, college and community issues
Question paper	Instructions and description of test tasks were evident; a lot of information supplied to fulfil task
Score sheet, scoring guide/ rating criteria	N/A
Instructions/guidelines	N/A

Commentary

Evidence from the test specifications showed a very brief description of the type of topics for the test. As with all other validity elements, a detailed description of each element is necessary for all parties concerned, especially since there are six parallel papers prepared for this speaking test every term, and not all this information was available in any document. Determining equivalent forms is a big task for the test developers; they need clear descriptors and guidelines on the choice of topics and the nature of information to be included in the test in order to produce valid and fair tests.

In general, information regarding test topic and content was missing or insufficient in the documents for the test; the question paper is the only evidence of the outcome of these descriptions and it cannot be a point of reference as topics change every year.

Table 4.20c

Observation	*Findings*
Researcher	Generally, students were not negatively affected by the topic and/or information in the test; they seemed to handle the information well and this was reflected in the presentations and group discussions

Commentary

Observations of various test sessions showed that most candidates were able to handle the topics and information found in the test papers rather well. They seemed comfortable with these topics and this was evident in that very few students had asked questions and for clarifications. This is not surprising in spite of the fact that students and lecturers felt they lacked background knowledge; interview data from both groups of participants revealed that students were exposed to a range of topics that they had to read up or research before the test.

Hence, while some students may lack background knowledge on the topics and test information, they would have benefited from the topics given to them in class and those which they had researched. Furthermore, data (table 4.20a) above indicated that the information provided in the test was useful in enabling students to generate ideas for their speeches. This however, contradicts the point that this aspect of test validity aims at measuring the test taker's content knowledge, in this case, especially background knowledge.

Language knowledge (definitions below are from Buck 2001: 104)

Linguistic knowledge on this side of the framework is meant to represent the model of language that drives the test – what the test developer assumes of the test taker [as opposed to the task demands side, in which it represents the operationalisation of the definition of language ability in the test task itself]. The aim is to investigate whether there is evidence that there is a coherent model driving the test; in this instance, the evidence should come from the developers and from the documentation.

Grammatical: Understanding short utterances on a literal semantic level. This includes phonology, stress, intonation, spoken vocabulary and spoken syntax.

Table 4.21a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 387)	Percentages higher for checking and adjusting grammatical accuracy in task B than task A; average 50% agreement
Questionnaire	Lecturer/Examiner (N= 46)	Percentages higher for checking and adjusting grammatical accuracy in task A than task B; all below 20% agreement
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	10 instances recorded: Language is a problem for many students; conscious of the errors & that low proficiency lowers confidence to speak
Interview	Lecturer/Examiner (N = 10)	Students struggle with word use & grammar; language ability depends on where they come from, rural or urban areas, east or west Malaysia, the faculty they are in
Interview	Administrator (N = 6)	Language usage is limited amongst the students, on and off campus
Interview	Expert (N = 5)	Language proficiency is a problem amongst all university students; the test attempts to raise awareness and help improve this situation

Commentary

Data shows lecturers/examiners having less confidence in the students' language ability, in fact this is the area where they usually lose marks in the test. Examiners were not certain if students were conscious of their word choices and grammatical accuracy because there is little evidence of self-correction. Students however, indicated that they are aware of the importance of 'good' English in order to speak well, and it is largely a matter of confidence.

Administrators and experts suggested that the speaking test, in fact, attempts to raise students' awareness of their language abilities so they are able to cope with the

demands of university courses. In addition, because English is a second, or in some cases foreign language for most students, their use of the language is limited. Given this evidence, it is then surprising that the students are still not able to cope with the linguistic demands of the test, since the administrators/expert would have made assumptions about students' abilities when developing the test. If data shows that students were not able to demonstrate their language knowledge when attempting the tasks, then there is a mismatch between the test developers' underlying assumptions regarding test takers' language knowledge and ability, and their actual levels of knowledge and ability.

In general, the evidence from this section of the study indicates that there is a broad feeling that students lack the grammatical knowledge that is required in order to fulfill the tasks and perform well in the test.

Table 4.21b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	No mention of grammatical knowledge required for the test
Question paper	N/A
Score sheet, scoring guide/ rating criteria	Rating criteria contain descriptors of three criteria: task fulfilment, language ability, communicative ability
Instructions/guidelines	N/A

Commentary

The only description of the language knowledge required to fulfill the test tasks are in the rating criteria which has brief descriptors of the three criteria used for rating. There is no evidence in the test specifications, question papers and the instructions/guidelines

for marking of the linguistic knowledge that students are expected to demonstrate in order to perform the tasks. A more detailed description of this validity element would assist lecturers to explain to the students details of the language knowledge required and assist examiners in their rating of the test, and perhaps prevent the mismatch mentioned in data above (table 4.21a).

Table 4.21c

Observation	*Findings*
Researcher	Students demonstrated some level of language knowledge; some were careful of word choice and sentence structures, especially in task B

Commentary

It was clear from the observations that there were students who were in control of their language use, especially in task A where they present their ideas individually; there was some evidence of stops and hesitations when errors in word choice or grammar were made, but not much on corrections. However, many others also struggled with their vocabulary and sentence structures here. It was also clear that in task B, the demands on grammatical knowledge were no less severe.

Discoursal: Understanding longer utterances or interactive discourse between two or more speakers. This includes knowledge of discourse features, such as cohesion foregrounding, rhetorical schemata and story grammars and knowledge of the structure of unplanned discourse.

Table 4.22a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	60% - 70% agreed they checked and adjusted the organization of their presentations 72% agreed they checked the effect of what they said on others during task B; 85% agreed they adjusted the points during the discussion
Questionnaire	Lecturer/Examiner (N= 46)	Low 20s% on students checked and adjusted the organization of their presentations Only 27% agreed students checked the effect of what they said on others during task B; but 82% agreed they adjusted the points during the discussion
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	It is difficult to make adjustments especially during the discussion because some speakers interrupt in the wrong places, agree too quickly and dwell on disagreements
Interview	Lecturer/Examiner (N = 10)	Students are able to present and conduct a discussion in the target language, but they lack the ability to maintain coherence at certain points in the discussion
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	For many students, presenting in front of an audience and participating in a discussion are not common activities; these are learned and performed in the classroom but rarely practised outside

Commentary

Although student data showed high percentages on this aspect of language knowledge, it is reasonable to assume that they actually had difficulty in checking and adjusting their speech during the tasks; this was mentioned by the students in the interviews. Data from staff confirmed that students were not able to adjust their language during the presentation and discussion because they rarely participate in these activities outside the classroom. Other factors such as first language interference affected their

ability to process the information at hand effectively. Expert data indicated an understanding of students' predicaments yet the discoursal functions here were not explicitly explained in any documents (table 4.22b below), or to the students in clear terms so they are able to attend to the task with recourse to knowledge of this element. Students' difficulty here is understandable; according to the definition, it requires the understanding of "discourse features, such as cohesion foregrounding and rhetorical schemata...", and these may be realized only with some knowledge of and extensive experience in using the language.

In general, discoursal knowledge appears to be a problem area for the students in this test.

Table 4.22b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	No description of discoursal knowledge in any of the documents
Question paper	
Score sheet, scoring guide/ Rating criteria	
Instructions/guidelines	

Commentary

Just like grammatical knowledge (table 4.21b above), there is no information in the documents on the discoursal knowledge required for these tasks.

Table 4.22c

Observation	*Findings*
Researcher	It was not clear whether students' performance were affected by lack of discoursal knowledge. There were, however, instances when students were incoherent in their presentations, this was clearer in the discussions

Commentary

Students showed instances of incoherence and an inability to respond appropriately to another speaker; in the second task, this affected the group performance and they were not able to conclude the discussion appropriately. As with evidence for grammatical knowledge above (table 4.21a), it appears that there is no clear sequence between what test administrators and experts expect of students in the test and students' actual demonstration of the linguistic function. Like data in 4.22a above, there is no clear link between what the test is testing and the students are able to do.

Functional: Understanding the function or illocutionary force of an utterance or longer text, and interpreting the intended meaning in terms of that. This includes understanding whether utterances are intended to convey ideas, manipulate, learn or are for creative expression, as well as understanding indirect speech acts and pragmatic implications.

The item in the questionnaire lists five functions of the speaking task: initiating a discussion, keeping a conversation going, connecting what one says to what has been said, taking turns appropriately, concluding a discussion.

Table 4.23a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	More than 70% agreement that they had no problems carrying out these functions
Questionnaire	Lecturer/Examiner (N= 46)	Percentages here **lower** than those recorded by students; the lowest is for 'connecting what one says to what's been said' (33% agreement)
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	14 instances recorded that it is important to: - Listen to other speakers - Not dwell on a point to be able to conclude - Take turns appropriately
Interview	Lecturer/Examiner (N = 10)	- Weak students speak without much thought to the functions; evidence of 'rehearsed' or 'learnt' phrases, expressions; minimal natural speech - Average students try to stay in the discussion & demonstrate one or two functions; some evidence of natural speech, some spontaneity - Good students are able to demonstrate any or all of these functions; speech is controlled and natural
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	

Commentary

Because of the nature of the tasks, the functions listed for this item in the questionnaire respond to both informational and interaction type tasks. Students indicated they were clear about the speech functions they needed to demonstrate and were aware of the importance of each one. Lecturers/examiners confirmed that students' functional knowledge is demonstrated within their language ability range; the weak, average and good speakers.

In general, students in this test are familiar with the speech functions that are required of them.

Table 4.23b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Functional range described in test objectives, e.g. stating and justifying opinions, accepting and rejecting ideas/proposals, and summarizing and concluding but not in detail of what each means and how students are expected to demonstrate them
Question paper	Evidence of this in the instructions and test task descriptions
Score sheet, scoring guide/Rating criteria	N/A
Instructions/guidelines	N/A

Commentary

There is evidence of the speech functions or functional range that students are expected to demonstrate in the test, though they are described as test objectives. An analysis of the documents showed that these 'functions' are listed but the level of functional knowledge and ability to demonstrate its use in the test are not clearly explained. Students need a clear description of the functional knowledge expected of them and how they can best demonstrate this in the test.

Table 4.23c

Observation	*Findings*
Researcher	Most students demonstrated functions for informational type task, such as stating a point and justifying, more than functions for interactive tasks such as conversational repair and meaning negotiation; this was either minimal or non existent

Commentary

Functions representing the construct of spoken ability (O'Sullivan, Saville & Weir 2002) are divided into informational (monologue) and interactive type tasks. The observations

showed most candidates demonstrating functions that are informational rather than interactional, even in task B where there is no evidence of agenda or discourse management.

Hence, while students have knowledge of the functions they need to demonstrate in the tasks, they were not able to demonstrate functions required of them to participate satisfactorily in an interactive task.

Sociolinguistic: Understanding the language of particular socio-cultural settings, and interpreting utterances in terms of the context of situation. This includes knowledge of appropriate linguistic forms, conventions characteristic of particular sociolinguistic groups, and the implications of their use, or non-use, in slang, idiomatic expressions, dialects, cultural references, figures of speech, levels of formality and registers'.

Table 4.24a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 387)	- Students **know** (83%) they have to speak differently when talking friends than when talking to staff/lecturers; they **use** (63%) different language when talking in English to friends than when talking in English to lecturers - 67% agreed they were able to conduct the discussion smoothly; 51% agreed they were able to conduct the discussion in an organized fashion
Questionnaire	Lecturer/Examiner (N= 46)	- 86% agreed students know this, 77% agreed students actually do this - Evaluation of student performance: below 50% agreed students are able to conduct the discussion smoothly and in an organized fashion
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 64)	Part of the performance has to do with whether you understand the other speakers, their ideas, language ability, etc. We know each other well, so we understand them quite well; it helps to improve the discussion.
Interview	Lecturer/Examiner (N = 10)	The dynamics between them in a group is important; they perform better with their classmates, when there is only one examiner present, when there is a leader in the group; personality within the group, i.e. shy, introverts do not contribute much
Interview	Administrator (N = 6)	
Interview	Expert (N = 5)	In general, students demonstrated the ability to interact & speak appropriately amongst their peers, though other factors such as personality, background etc. impede their performance

Commentary

It is interesting that students and staff commented on the importance of personalities and abilities within a group, which highlights the consequence of the group interaction task and how students are grouped together. This is one aspect of sociolinguistic knowledge; students are aware to a certain extent of how their friends think and behave in a group and that this affects performance in the discussion. Lecturers/examiners

suggest the importance of group dynamics, such as the shy, introverted student will usually contribute minimally to the interaction. (Berry 1993,1994, 1996)

In general, most participants show awareness of the importance of sociolinguistic knowledge relevant for the tasks in this test; this is true especially because they know the other speakers in the group and may understand them better. Understanding sociolinguistic characteristic of other speakers will affect behaviours and cognitive processing in the test.

Table 4.24b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	No description of sociolinguistic characteristics and knowledge required in the test in any of the documents
Question paper	
Score sheet, scoring guide/Rating criteria	
Instructions/guidelines	

Commentary

Like all other language knowledge elements, having sociolinguistic knowledge of the test context, participants involved, and so on would be advantageous for the test taker as this knowledge will enable him/her to process language use efficiently; this was confirmed by interview data from other participants/staff (table tables 4.21a, 4.22a, 4.23a above).

Table 4.24c

Observation	*Findings*
Researcher	Some students seemed to demonstrate this knowledge as they were comfortable communicating and interacting with other speakers in the group; some evidence of filling in the pauses, helping out those who were not able to continue & a leader within a group

Commentary

It was apparent that students had some level of understanding of group dynamics or composition; they were able to control levels of formality, and understood cues from each other, when to speak or not and the order in which they should speak. These conventions were probably taught to them or experienced by them in the English class, and throughout the term presenting speeches and even interacting with their classmates only.

4.3.7 OVERALL COMMENT ON EXECUTIVE RESOURCES

As stated at the beginning of this section, the aim here is to investigate whether there is evidence on assumptions that test developers, writers, experts etc. make about students' ability and knowledge of spoken language which are required in order to attend to the test tasks. The evidence for this is clear in terms of documentations and data from test administrators and experts.

The documents do not contain clear descriptions and details of the elements' content and language knowledge, which the test developers expect of the candidates. There is an abundance of information that needs to be included here and taught to students, such as what it means to be able to initiate a discussion or manage turn taking

(functional language), and having knowledge of interactive discourse features such as knowledge of the structure of unplanned discourse. However, there was no information on what the elements are in terms of candidates' cognitive processing and how best to demonstrate them in the test. While administrators and experts commented on candidates' inabilities and factors that affect their performance in the test, we constantly find evidence from student data that contradicts this testimony.

Hence, we find a mismatch in terms of what test developers expect of the candidates, reflected in the test instructions and task descriptions, and what the candidates' true ability and knowledge are in terms of these expectations. Moreover, this mismatch reflects the absence of a theoretically driven test; test administrators and experts were not able to express the concerns in terms of a theory/model of language ability on which the various aspects of the test were developed.

Students did employ certain strategies in attempting the tasks, but their performance did not reflect the intensity involved in processing (mostly at preparation time, and occasionally at presentation/speaking time) and did not show the more salient functions of negotiating meaning, indicating understanding, conversational repair, and so on. There are possibly two explanations for this. Firstly, students had some practice in class to express the functions above, in line with the requirements listed in the syllabus. Secondly, unlike ordinary conversation, it is probably not possible for candidates to demonstrate many of the functions in a testing situation, especially when language, time, anxiety and other task demands are factors that interfere rather than facilitate performance.

More importantly, candidates are often disadvantaged because they are not given explicit instructions and knowledge on how to process the task demands, given their own resources in terms of content and language knowledge.

4.3.8 CONCLUSIONS FROM THEORY-BASED VALIDITY ANALYSIS

In conclusion, it can be argued that the test lacked theory-based validity and this is related to lack of clarity in terms of what test developers and administrators hypothesize the students' language knowledge and abilities to be and how this is documented in the test. Test developers would usually have assumptions about the candidates' abilities based on their own data from research conducted during the test development process, and/or a theory or model of learners' language ability with respect to cognitive processing of speech in the context of the direct test above (see Bygate 1987, Levelt 1993, Hughes 2002 for speech processing in context). These assumptions are then translated into specifications for the test describing resources and processes required to prepare for and perform the tasks. If these are not clearly expressed in test documents for the administrators, raters and students, a discrepancy is found between test demands and student performance.

The candidates in the test seemed to have some awareness regarding the processes they underwent during the test, such as monitoring their speech which is partly influenced by co-construction of discourse, but failed to understand the various linguistic knowledge required to fulfill the tasks efficiently, mainly due to lack of explicit information available to them in the test documents.

In principal, the evidence seems to be overwhelming that the test lacks a clearly defined model of spoken language ability on which it was developed; this makes the test essentially atheoretical.

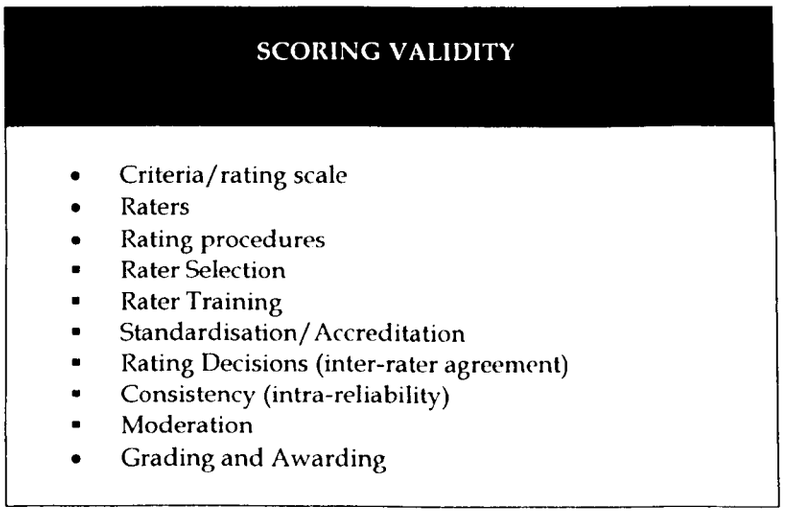
4.3.9 SCORING VALIDITY

Scoring validity evidence concerns candidates' test scores: how rating conditions and all factors related to the rating procedure affect the reliability of test scores assigned to the candidates. In this aspect of validity, it is the identification and minimization of such errors of measurement, actual or potential, which must concern the test developers (Ellerton 1997), so that the procedures for assigning test scores, and the test scores themselves are reliable.

The elements of context validity are considered at the test design stage because it can potentially impact the reliability of the test. Hence, just as context validity affects theory-based validity, it can also influence test reliability, i.e. how test developers determine the criteria for rating student performance is directly related to the test tasks. The operations that we expect students to perform under various conditions of the test determine the type and number of criteria for rating the tasks.

Figure 4.6 below (Weir 2004) shows the elements of scoring validity, which are reported according to data gathered in the present study from all participants.
The students **did not** complete the section in the questionnaire on 'Scoring validity'; however, they referred to the rating process in the interviews.

Figure 4.6 Aspects of Scoring Validity for Speaking (from Weir, 2004)



Criteria/Rating scale Appropriate criteria for assessing speaking have to be agreed upon at the test design stage and they need to reflect the features of spoken language interaction the test task is designed to generate. Tasks cannot be considered separately from the criteria that might be applied to the performances they result in (Weir 2005, p. 192). The consideration then is how best to apply the criteria to the samples of task performance.

The UiTM rating scale has three criteria: Language ability; Communicative ability; Task fulfilment.

Table 4.25a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Lecturer/Examiner (N= 46)	73% agreed three criteria cover all aspects of task performance, & they are sufficient for a fair judgement; 54% agreed they are clear to all markers; **May** data: 80% of participants agreed that criteria were unclear for examiners
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Lecturer/Examiner (N = 10)	- Descriptors are brief and vague ('excellent' user,' good' user, 'limited' user, and so on) - Raters are interpreting each criterion differently; e.g. Communicative ability/CA not clear to most examiners; needs to be revised, different one for each task, examiners interpret it in many ways, e.g. could be fluency, non-verbal cues, confidence, etc.
Interview	Administrator (N = 6)	No formal training on use of the scale but a session for speaking test lecturers/examiners at the start of term on issues to raise & discuss
Interview	Expert (N = 5)	Some lecturers have had experience being rater for the MUET; they would be trained by the Exam Council

Commentary

Lecturers and other members of staff agreed that while the three criteria were sufficient and appropriate for the tasks, they stated that the descriptors of these criteria were brief and vague. There are three similar components, score points (1-6 for TF,LA, 0.5-3.0 for CA), and very brief descriptions of what each score would indicate for both tasks. For example, a 4 on TF is "fulfills task satisfactorily", on LA is "displays satisfactory control of language", and on CA this is equivalent to 2.0 "shows ability to communicate satisfactorily". Other expressions/adjectives used to describe performance include "competently", "modestly", and "limited" or "poor" control of language or content.

Further data, collected in May 2005 showed more staff participants stating the same point: the criteria are not clear enough for examiners to make their judgments on test

performance efficiently and fairly. Administrators added that it is worrying that raters are interpreting each criterion differently, and agreed that they need to be revised. This is evident in data on rating the test using the existing scale (see Appendix 3.10); the range in the scores given were wide, showing how the examiners are marking the same test for the same candidates, using the same scale, but in different ways. This clearly shows that no clear model of language ability lies behind the test. Because of this nobody really knows what is being tested and control over the test process is lost to the tester, and the results are likely to be essentially meaningless.

Thus, the criteria for rating this speaking test are not explicit enough for the examiners to rate the performances efficiently.

Table 4.25b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	N/A
Question paper	N/A
Score sheet, scoring guide/ Rating criteria	Each criterion described, accompanied by a rating scale in the scoring guide/rating scale: Task fulfilment (range 1-6); Language ability (range 1-6); Communicative ability (range 1-3)
Instructions/guidelines	A guideline for scoring mentions confidentiality of the assessment process & recording of marks.

Commentary

The descriptions of the three criteria for rating the test and guidelines for scoring are found in a set of documents for scoring the test. It was evident that each criterion had a set of descriptors according to the range of possible marks a candidate can attain, e.g. task fulfillment (range 1-6), '6 = Fulfils task very competently', '4 = Fulfils task

satisfactorily', and there is a separate scoring guide for each task, even though the criteria are the same for both. The descriptors seemed vague as they range according to the scores from the highest score (6; 3 for communicative ability) which is described as 'very competent' to the lowest score (1; 0.5 for communicative ability) which is described as having limited skill/ability or none at all.

There were no other documents found relating to the criteria. The document above contained information that was not helpful and detailed for the raters to be able to use effectively.

Table 4.25c

Observation	*Findings*
Researcher	During the tests, examiners appeared familiar with the documents and had no problems using them.

Commentary

Most examiners were familiar with the documents for rating the test because they had conducted the test several times before, and the usage is explained in the guidelines for marking document, such as examiners are required to compare their marks after each test to check for differences and errors in rating. However, the interview data above indicated that this information is minimal and insufficient for raters to realize the vital features of the criteria for rating both tasks.

Rater training A systematic process to train raters to apply the rating scale and the mark scheme in a consistent way. This involves careful consideration of the context in which

training occurs, the type of training given, the extent to which training is monitored and feedback given to raters.

In the UiTM speaking test, as with other tests, the questions on rater training are:

Does it happen? If not, why?

If so, who is responsible?

Who does it? How is it conducted? Was it successful? Were raters satisfied?

Table 4.26a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Lecturer/Examiner (N= 46)	Low percentage of agreement on adequate information on the rating process (55%) and rater training (41%); **May** data: very low on rater training (26%)
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Lecturer/Examiner (N = 10)	Training is insufficient; the main campus conducts more sessions than the branch campuses
Interview	Administrator (N = 6)	Persons-in-charge give briefings on the assessment process at the beginning of term; formal training provided by the Examination council
Interview	Expert (N = 5)	Formal training not given to lecturers/examiners or not organised consistently, especially in branch campuses

Commentary

Data indicated that lecturers/examiners receive insufficient and sometimes informal training, and inadequate or insufficient information on the rating process. Though administrators said that the persons-in-charge of the speaking component are responsible for this and that they do this at the beginning of each term, the lecturers/examiners stated otherwise; they lack formal training & adequate information. At the university, differences in terms of frequency and type of training provided are of

special concern; lecturers at the main campus receive more training than those at the branch campuses, and some of these lecturers have the opportunity to take part in more formal training sessions conducted by the Examination Council. The experts agreed this is an area of concern and that the university needs to look into urgently. Luoma (2004) stressed that raters have to make practical decisions during the rating process, such as rating task by task or the use of certain criteria for one task and not the other; if they are not well-trained for the task, marker reliability and validity of the test as a whole will be affected.

Hence, lack of rater training and adequate information on the rating process are problems the lecturers face in this test, especially since one is an examiner/rater and she or he also manages the test, and the other is a rater only. In fact, the answer to the questions asked above is clear: There is no training conducted consistently for the lecturers. No reason is provided for this, the situation is unsatisfactory and most lecturers are unhappy about unfairness and inadequacy of training across campuses.

Table 4.26b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	No information available here
Question paper	
Score sheet, scoring guide/ criteria	
Instructions/guidelines	Describes examiners' duties briefly; on taking the leading role, and moderation of marks after the test

Commentary

The only available information is regarding the examiner's duties: one of the examiners who is also a rater, takes the leading role and manages the conduct of the test and moderation of marks after the test. There is no mention of rater or examiner training in any of the documents.

Table 4.26c

Observation	*Findings*
Researcher	Examiners rate the performances task by task (task A followed by task B); they seemed confident of conducting the whole assessment process. Unable to conclude if they were marking systematically and objectively

Commentary

It was observed that no problems were raised during the tests regarding the rating process, even though it is difficult to see if raters/examiners were marking systematically, and data from the questionnaires and interviews indicated that there were problems. Like the use of the rating criteria/scale, these lecturers/examiners are familiar with the rating process from experience at the university, and some had experience as examiners and as raters for the national MUET speaking exam. Hence, though the lecturers had no problems in the rating process, they realize the importance of proper training and knowledge on rating.

Standardization The process of ensuring that markers adhere to an agreed procedure and apply rating scales in an appropriate way. This is to bring examiners into line, so that candidate's marks are affected as little as possible by the particular examiner who assesses them. Standardization is thus included as a part of the rater training process.

Table 4.27a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Lecturer/Examiner (N= 46)	64% agreed that raters are standardized to benchmark candidate performance levels before marking/rating begins
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Lecturer/Examiner (N = 10)	Standardization process inconsistent each term; some times none at all
Interview	Administrator (N = 6)	In spite of standardization, there are inconsistencies in marking, especially across campuses
Interview	Expert (N = 5)	Disparities in that most examiners do impression/ holistic marking first, then look at the criteria, and inconsistencies in that some are too strict, and others too lenient, while some use own benchmarking to mark the group, and some others stress on one criterion over others

Commentary

Lecturers pointed out that like training, the standardization process is inconsistent in terms of frequency and also method. These issues were raised by the person-in-charge of the speaking team who expressed her worry about the data she has on the wide disparity between scores awarded by examiners at the branch campuses and examiners at the main campus. One of the major problems is whether examiners are *standardized to benchmark candidate performance levels before marking/rating begins* (stress from Weir 2005). In the Malaysian context, some lecturers who have had experience as a MUET (Malaysian University English Test) examiner and/or rater would have received formal training from the Examination Council. Still, many lecturers have difficulties with awarding marks between bands 3-4 ('modest' to 'competent' user of the language),

which according to several experts (Head of speaking test team, Academic coordinator, Resource person Mainstream English 2, Head of English) is the range in which a majority of the students fall into.

It is obvious that inconsistencies in marking may have resulted from lecturers who do not have a clear understanding of the standardization process.

Table 4.27b

Document Analysis	
Document	***Findings***
Syllabus/BEL250/Test Specs	No information available regarding standardization of examiners/raters in these documents
Question paper	
Score sheet, scoring guide/ Rating criteria	
Instructions/guidelines	

Commentary

There is no mention of the standardization process in any of the documents.

Rating conditions under which marking takes place, e.g. temporal, physical, and psychological, have a potential impact on scoring and need to be standardized too.

Table 4.28a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Lecturer/Examiner (N= 46)	68% agreed that raters are able to work without any disturbance or distraction during the rating process; 59% agreed two lecturers are present at the tests
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Lecturer/Examiner (N = 10)	Problems in rating conditions due to shortage of rooms and examiners, and time for the test, affect the reliability of the rating process and the validity of the test as a whole; this is especially true with unfavorable test conditions in the branch campuses
Interview	Administrator (N = 6)	Coordinators and lecturers at branch campuses organize their own conditions for the test; problems above especially serious in branch campuses
Interview	Expert (N = 5)	Problems in rating vary between campuses; many have only one examiner present due to logistical constraints + one examiner is class lecturer ; these affect reliability of test scores when a lecturer marks his/her student in terms of their competency and not performance during the test. Many examiners also uncertain about rating the bands 3-4 candidates

Commentary

It is clear from the interview data that rating conditions and rating processes vary between faculties in the main campus, and between campuses. The branch campuses seem to suffer most from unfavourable and insufficient test conditions, such as shortage of rooms, examiners, time constraint, and so on. A major problem in the rating of this test is the presence of a class lecturer of the group of candidates as a second examiner. The bias of knowing one's students' abilities influences marking, i.e. as the experts pointed out, they are rating based on their knowledge of a student's competency rather than the actual performance at the test; this can result in either positive or negative bias.

Another problem is the presence of only one rater at many centres especially in the branch campuses.

Rating is affected by all these factors considerably, and data from all participants confirmed the point: marker reliability and reliability of test scores are affected and this affects test validity as a whole.

Table 4.28b

Observation	*Findings*
Researcher	Of four branch campuses visited, one had reasonable conditions under which the assessment took place; many also had only one examiner present, this was observed in the main campus as well

Commentary

The observations confirmed what all participants claimed in the interviews, that rating conditions were more favorable in some campuses than others. These include physical conditions, such as tests conducted in lecturers' office, and time considerations that resulted in shortage of examiners and unsuitable test times. In addition, because one of the examiners, or in some cases the only examiner present, is the class lecturer, tests seemed to run smoothly as the planning had been carried out between lecturer and students; however, this does not ensure the reliability of marking and test scores, and often it doesn't. Inadequate conditions and conditions which vary too much across campuses cause disparity in marking and unreliable scoring.

Moderation of scores The advantages of a double as against a single marker system would only be clear if the two markers are equally consistent in their own marking; if

this is not the case, reliability of a more consistent single marker would be better than the combined reliability estimate of the two inconsistent markers. However, as mentioned above, an adequate marking scheme (table 4.25) and sufficient standardization of examiners (table 4.27) would ensure a high standard of inter-rater and intra-rater reliability.

Table 4.29a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Lecturer/Examiner (N= 46)	59% agreed that marks are moderated after the test to sort out any differences or problems between the raters
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Lecturer/Examiner (N = 10)	Raters add up their marks and average those to arrive at a final score; differences or errors addressed here. After many years, students are obtaining between 8 to 10 marks, i.e. bands 3-4
Interview	Administrator (N = 6)	Examiners are required to moderate the marks after the test; this is stipulated in the guidelines for marking
Interview	Expert (N = 5)	Candidates are moving closer to the target band by the year, examiners are becoming more experienced, but the differences in marking are still there

Commentary

Moderation does not appear to be a problem for this test when there are two examiners/ raters present to compare their marks after the test to check for big differences and errors. But when there is only a single examiner present, as data above (table 4.28a) indicates, marker reliability is a major problem. Even though the administrators confirmed that this is stipulated in the guidelines for marking the test, they are also

aware of examiners who do not comply; this is evident at the end of the assessment when random checks are conducted on the score sheets and discrepancies are found. Hence, there are still cases of inconsistent markings and evidence of no moderation between examiners, but those who comply with this regulation to ascertain consistency in their marks at the end of the assessment process do so through moderation with the second examiner/rater.

Statistical analyses on candidates' test scores.

Table 4.30a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Lecturer/Examiner (N= 46)	Only 18% agreement on whether statistical analyses are conducted on the marks to check consistency and level of marking
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Lecturer/Examiner (N = 10)	Very few lecturers take initiative to analyze students' test results; marks are moderated & sent to the main campus for further analyses. Mixed response on whether the scores between different components correlate, i.e. if they do well in speaking, they would do well in other skills too
Interview	Administrator (N = 6)	Plenty of data from previous exams but little analyses on the speaking test scores, on rating and performance; lately receiving students' scores from the MUET so would be good to analyse these results to see which components they are doing well in
Interview	Expert (N = 5)	Examination council not transparent enough on analyses done on MUET test scores; not surprising UiTM not doing the analyses; the university needs to analyze how their students are performing in the tests

Commentary

It is clear that most lecturers/examiners and the administrators do not conduct further analyses on the speaking test results, in terms of marker reliability and student

performance. Most staff participants had mixed ideas about the relationship between the skills and whether they can actually be compared; some stated that data seemed to show correlation between speaking and writing. i.e. a student with high speaking marks tends to receive fairly high marks for writing, usually lower than the speaking marks. It is interesting to note that experts highlighted the transparency of test scores by the Examination council; this practice however, does not rule out the analyses conducted by the university which can only benefit members of staff and students in general. Consequently, administrators could perform further analysis by using multi-faceted Rasch or a classical analysis correlation coefficient to provide them with a clearer picture of whether examiners are behaving consistently with themselves and with other markers, and even detect any bias against individual candidates.

Table 4.30b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	No information relating to statistical analysis of test scores in all documents
Question paper	
Score sheet, scoring guide/ Rating criteria	
Instructions/guidelines	

Commentary

There were no evidence of statistical analyses in the documents above, but there was evidence of some analyses conducted on the results by the testing team for the exam meeting. Analyses were in the form of an overall description of how students performed in the course in terms of descriptive statistics and grade received; these were found in

the file of score sheets compiled by the academic coordinator and analyses prepared by the testing committee members.

Grading and awarding This is the final part of the scoring process where grades are decided and checks carried out to ensure the test was not biased against any group of candidates.

Table 4.31a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Lecturer/Examiner (N= 46)	Only 27% agreement on whether all results are checked by the exam committee to ensure fairness
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Administrator (N = 6)	The exam committee members, coordinators and resource persons for each course check the results together for accuracy and so on, before the results are sent out
Interview	Expert (N = 5)	

Commentary

Data on this aspect of the test is scarce even though administrators confirmed that they conduct a grading process (see Weir & Milanovic (2003) for details) at the end of the assessment period every term. At this meeting the cut off score for the various grades are set, and other reports and analyses that have been carried out on score data are reviewed before the results in terms of grades are generated.

Unfortunately, this process is not formally recorded in the documentation, and therefore may not be conducted in a similar fashion for every administration.

Table 4.31b

Document Analysis	
Document	*Findings*
Syllabus/BEL250/Test Specs	Information not available in any of the documents here
Question paper	
Score sheet, scoring guide/ Rating criteria	
Instructions/guidelines	

There is no clear evidence of grading and awarding in the documents above.

4.4 OVERALL COMMENT ON SCORING VALIDITY

We can conclude that for scoring validity, only two elements were favourable in the test: moderation and grading/ awarding of marks, although information here was not documented systematically. For all the other elements, the test failed to provide sufficient evidence to justify any validity argument. There were serious concerns raised for the criteria/rating scale, rater training, standardization of examiners, the rating process and rating conditions, and statistical analysis on test scores.

Inconsistencies and variability in marking the speaking test above can be attributed to a number of factors. These include: inadequate rating scale/criteria for rating, rater expertise and training, and inadequate rating conditions and rating process. One useful suggestion to overcome these concerns is certainly the use of video cameras to record the candidates performing the test activities, and these can subsequently be used for training and standardization of marking (Weir 2004). With regard to rater

inconsistencies, the use of statistical tools such as the multi-faceted Rasch (MFR) model (Linacre 1989) can provide a clear picture of whether markers are behaving consistently within themselves and with other markers; it can also detect any bias for or against individual candidates. In addition, MFR analysis allows investigation of the influence of other 'facets' within the assessment procedure that can also impinge on the outcome; the influence of tasks compared with other facets such as the rater and the rating scale.

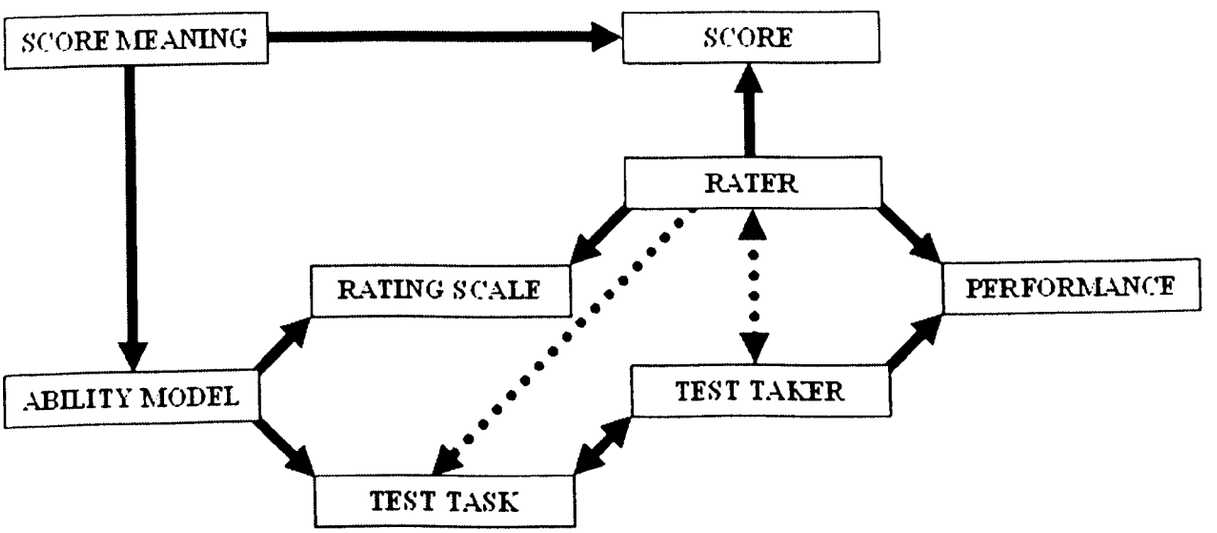
A pressing issue related to this is the grading or assessment of interactive tasks such as the group discussion in this speaking test. Research on using group interaction tasks in testing speaking seems divided. Like pair work, it is well received by learners, but is problematic because of administrative concerns about managing the size of the groups and the mixture of ability levels in them (Luoma 2004). It is also more suitable for classroom assessment rather than in formal tests of speaking (see Reeves 1991). In addition to this, rating group interaction is a difficult task due to several factors which are listed in this chapter and were discussed in detail in chapter 2. Mainly, a group consists of candidates with a mixture of ability levels and the task of the rater is to assign each candidate with an individual score which is based on a group performance.

Recent studies have alerted us to the caveats associated with paired or group tasks in testing speaking, which causes the marker great difficulty and testers the problem of interpreting individual scores based on a jointly constructed interaction among participants (see Dimitrova-Galaczi 2004, for details on issues of interaction in paired speaking tests).

The findings above (tables 4.26, 4.27, 4.29) relating to poor or inconsistent training, standardization and moderation are evidence of such difficulties, especially when these elements of validities are deficient in the test. Moreover, if these are not overtly linked to the model of ability that drives the test, they are meaningless, and since there is no model, the test outcomes in terms of test scores are essentially meaningless.

Figure 4.7 below (O'Sullivan 2005) shows how the underlying model or theory of ability which steers the test is directly linked to test tasks and the rating scale, which themselves affect rating, performance and the eventual meaning that we make of the test scores.

Figure 4.7 Model of the test process



This is the key consideration for test design and development, even before all other aspects of the test are considered and decided upon. In the study on the UiTM speaking

test, we found discrepancies in test administration and rating, insufficient and inadequate information in test documentation, while participants raised concerns and misgivings regarding many of the validity elements, mainly because they were unclear about the essential elements of the test.

4.5 SUMMARY OF VALIDITY OF THE DIRECT SPEAKING TEST

We could say that the whole test is based on the tasks. It is not clear what theoretical basis there is for these tasks. The scoring criteria are also based on the tasks, and not on any underlying model. This means that each test is essentially unique, and comparisons across different versions are meaningless. The model anchors the different versions of the test; without a model, there is no anchor, hence no explicit relationship.

Without a model there is no meaning, and without meaning there is no validity. Hence we conclude that without any of the elements here we have no convincing evidence of validity for the test.

Figure 4.8below summarizes the findings relating to context validity, theory-based validity and scoring validity of the UiTM speaking test.

Figure 4.8 Summary of validity of the direct speaking test

Validity Component	Validity Element	Positive finding	Negative finding	Mixed finding
CONTEXT VALIDITY	Purpose	♦		
	Response format			♦
	Weighting		♦	
	Known criteria			♦
	Order of items		♦	
	Time constraint		♦	
	Physical conditions		♦	
	Uniformity of administration		♦	
	Security			♦
	Channel		♦	
	Discourse mode		♦	
	Test length		♦	
	Nature of information		♦	
	Content knowledge required	♦		
	Lexical range		♦	
	Structural range		♦	
	Functional range		♦	
	Interlocutor variables		♦	
THEORY-BASED VALIDITY	Conceptualiser			♦
	Linguistic formulator			♦
	Overt Speech			♦
	Internal knowledge		♦	
	External knowledge		♦	
	Grammatical knowledge		♦	
	Discoursal knowledge			♦
	Functional knowledge		♦	
	Sociolinguistic knowledge			♦
SCORING VALIDITY	Criteria/Rating scale		♦	
	Rater training		♦	
	Standardization		♦	
	Rating condition		♦	
	Moderation			♦
	Statistical analysis		♦	
	Grading & awarding			♦

It is clear from the table that the test lacks validity especially in the areas of context and scoring validity. The data is for the most part mixed in theory-based validity; students are disadvantaged in demonstrating their language knowledge and abilities because of the shortcomings of the test in terms of validity of test design and development, its documentations, and reliability of the rating process.

Chapter 5: THE SEMI-DIRECT COMPUTER TEST

5.1 INTRODUCTION

This chapter outlines the underlying principles governing the development of the computer test, the operational model for its development and operations, and findings on the validity of the test.

The rationale for the development of a semi-direct web delivered speaking test was established in Chapter 1. As was stated, the main reason for this new channel of testing speaking arose from the needs of Mara University of Technology, Malaysia, where the speaking test is administered as a direct, face-to-face test to over 10,000 full-time and part-time students at various locations across the country every year. The test takes many hours of developing, administering, rating, and analyzing by members of staff at the Language Academy and the university as a whole. The major concern here is how conducting the test on such a large scale affects efficiency, reliability and validity.

In addition, research on computer-based testing has to date focused on testing of reading, writing or grammar, with research on testing of speaking, especially in the semi-direct mode, being limited and less popular. In fact, actual work that is in progress on computer-delivered speaking tests have been adaptive (CAT) in nature and conducted on monologue type presentations. Examples of these are found mainly in Kenyon & Malabonga (1999- present); TOEFL CBT or the new speaking test (TOEFL Academic Speaking Test/TAST, see http://www.ets.org); more recent research on

computer-based IELTS (see Cambridge ESOL site at http://www.cambridgeesol.org). Referring to chapter 2 in which the literature on computerized testing was reviewed, it was clear that research on testing spoken language using the computer is scarce, yet making progress, even if at a slower pace than testing of other skills. In terms of the current study, we hope to draw on the best of the direct test of speaking in terms of its theory-based and context validity, and semi-direct test in terms of reliability and practicality to develop a speaking test that may meet the necessary requirements of validity, reliability and efficiency. However, several considerations are made in this attempt to deliver the speaking test using computer technology.

Firstly, we emphasise that the purpose of the study is **not** to develop a new computer-based speaking test, but to explore issues relating to the direct test through the validation process. In fact, the computer test was developed from a revision of the existing UiTM speaking test after it had been validated (see Chapter 4: The Direct test Validation), and all aspects/features of the old test were retained for the computer test, i.e. the topic(s), format, thematic link between tasks, instructions, and so on. Indeed, the only major difference is in the mode of delivery, i.e. using the computer, even though suggestions were made for a new set of rating criteria/scale for the computer test. The main reason for doing so is for comparison purposes, see point three below.

Secondly, speaking tests usually consist of single or multiple tasks (interview, presentation, discussion); for the CB test, tasks in the old test were retained, i.e. an individual presentation, followed by a discussion (though interaction here is between the candidate and 'non-live' interlocutors). In addition, the language that candidates

produced in the old test was established to ensure that similar functions are elicited by candidates in the computer test. After viewing several video clips of the direct test using the checklist for speech functions (O'Sullivan, Weir & Saville 2002) to determine language produced by the examinees for both tasks, it was found that they were mainly informational in nature, with minimal or no interactional functions, and no agenda (or discourse) management. Subsequent viewings of other candidates doing the same test revealed similar results. Hence, this aspect of the test was a major consideration in the development of the computer test in terms of the language functions that the candidates are expected to produce (the development of the computer test is detailed later in this chapter).

A third point refers to parallelism in testing; since the study is based on the speaking test conducted at the university, the computer test needs to parallel the existing test in as many aspects as possible (this is in order to facilitate comparison). This is a key consideration in order to enable a comparative analysis of the two modes of testing speaking at the end of the study, which will then address the final research question, which relates to the possibility of replacing the direct test with the web-based one. Moreover, it was highlighted in chapter 4 that one of the issues raised by staff and students at the university is one of equivalent forms of the test. To ensure security and confidentiality, six parallel forms are developed each year and distributed to campuses across the country, and specific instructions about the order in which they are to be used are provided. However, data from the validation study showed that ensuring security is still a major problem amongst test administrators; test takers found the topics/task situations imbalanced and that some were more difficult, between the test sets, as well

as within one test set. As we explore the use of a computer- delivered test, it was clear (this will be discussed later in this chapter) that the issue of equivalent forms does not surface in the computer test as all candidates receive the same input in terms of instructions and task situations.

5.2 UNDERLYING PRINCIPLES

The semi-direct computer version of the speaking test was developed based on various principles. First, the literature on computerized testing which included computer-based, computer-adaptive, and web-based testing were reviewed and revealed that little research on computerized speaking tests exists to date. As detailed in chapter 3 and as stated above, the idea for a computer-delivered test was based on the rationale for its development, and the objectives of the study, i.e. to operationalize a framework for validating speaking tests (Weir 2004/2005), which would enable us to provide an evidential basis for replacing a direct test of speaking ability with a semi direct web-based speaking test. Hence, the framework was operationalized in the form of questionnaires for the various validity components, and these were used as an instrument for gathering data on test validity.

Secondly, this stage of the study refers to the second research question that developed from the literature and objectives of the study:

To what extent is a proposed semi direct web-based speaking test valid in terms of:

a) *content validity*

b) *theory-based validity*

c) scoring validity

In addition, data gathered from the direct test validation study (chapter 4) were closely considered through the entire development of the computer test. Hence, the third principle points to validation. The importance of validation has resonated throughout this study from chapter 1; the question we ask is not an 'all or none' question, but a matter of degree of validity that a test achieves (Messick 1989). Weir (2005) emphasizes that claims of validity can easily be made but many examinations often lack validation studies of actual tests to demonstrate this. Like the direct test validation study (Main Study 1), the computer test is also subjected to a validation study (details of findings detailed later in this chapter) to ascertain its validity in terms of the purpose for which it was developed. The validation exercise was also based on the same socio-cognitive framework for validating a speaking test. Data for this validation study was gathered from candidates who participated in the computer test at various stages, and staff members who were shown the test in several workshop sessions (in Main Study 2/Feb-Mar 2005 and May 2005) and during the computer test trials.

A final important note is related to the main aim of the study: it is not to develop a new web-based speaking test, but rather to replicate the old test using a new or different platform so that problems that rose from the direct test may be addressed. One way of achieving this goal is by conducting a validation study in a systematic and organized manner, using a framework for validation, to ascertain the problems of the test. We then deliver the test on the computer to determine if the issues and problems can indeed be resolved using this method.

The computer test was therefore, built on the format of the existing test and data gathered from the validation exercise to make them parallel. Only then will we be able to make comparisons between them in terms of reliability, validity and efficiency, in order to address the research questions of the study.

5.2.1 Limitations of the direct speaking test

Based on data gathered from the direct test validation study (chapter 4), the following shortcomings of the test were discovered; these are summarized below.

1. A major shortcoming of the direct test was in its design. The test was based on a series of tasks, but how they were determined was not evident ; this had a serious effect on participants' (test takers and staff) understanding of key elements of the test which students could capitalize on to perform well in the test. In essence, it was not designed or driven by a theory or model of language ability and this in turn affected the elements of context, theory-based and scoring validities. (details in number 2 - 4 below; see chapter 4, figure 4.5 for a summary)
2. The test lacked context validity in terms of all elements except test purpose, lexical range and some interlocutor variables. Test settings such as criteria for rating, time constraints and weighting for each tasks were not made explicit to candidates in the test paper. Variation in physical conditions across campuses and inconsistencies in test administrations affected test conditions and performance. Candidates and staff members lacked knowledge and clear understanding of task demands such as nature of information, content

knowledge required, structural and functional range; in turn, candidates were not able to demonstrate their best abilities.

3. Lack of understanding of context-based validity elements above led to serious problems in theory-based validity. Candidates were unable to process the information efficiently because they lacked knowledge of strategies and resources needed in the test. Lecturers were not able to make clear to students what cognitive processing involves, and how they are expected to express this in the test.
4. Scoring validity is a major concern of the test. Examiners and raters lacked training and standardization, were not equipped with a comprehensive and lucid set of criteria and rating scale, were burdened with test administration and unreasonable conditions, all of which affected test reliability.

The significance of the results of the existing speaking test is clear: not only are we able to ascertain the validity consequences of the test in detail, but also realize the symbiotic relationship between the validity components. How we design the test and what elements we include in it will have a direct impact on the internal processing that test takers set in motion in order to attempt the test tasks. This in turn influences how examiners and raters mark a candidate's performance, using a set of criteria and/or scale that was developed based on the characteristics of the test tasks. Each component complements the other in this relationship; if the test falls short of an explicit design, based on each validity element of the framework, test validity becomes deficient. In these terms, though our computer test replicated the old test, we had taken into consideration and made the changes in terms of the test setting and task demands, and

even proposed a new rating scale. Thus, the delivery of the test using computer technology could directly address the concerns which were raised in the research of the direct test.

The rest of this chapter focuses on the web-based test: its development, operations, and data collected at various stages to establish its validity as an instrument for delivering the speaking test.

5.3 MODEL OF THE TEST DESIGN AND OPERATIONS

The development of the computer test was described in detail in chapter 3 (see section 'Instruments for Main Study 2'). However, we refer to diagram below (Figure 5.2) which illustrates the various stages of test development and the processes involved in gathering data for its operations and validation.

5.3.1 Test design phase

The computer test was designed based on the existing UiTM speaking test, and the stages involved in its design are indicated in Figure 5.2 below (see boxes in blue), and described as follows.

A. The monologue test (task A)

The test was designed based on a series of speaking test topics, designed to elicit monologic discourse forms and researched by colleagues at CLARe (Centre for Language Assessment and Research), Roehampton University. As indicated below, the test had been adopted and trialled at the university with EFL

students; the task of the researcher for the current study was to conduct further trials of the task type in Malaysia and then to devise versions related to the specific context.

B. The interaction test (task B)

The design for task B entailed many stages, as this was the first time it would be developed; unlike task A above, there were no samples or models to replicate. The preliminary trials were conducted at Roehampton University before it was finalized and conducted in Malaysia.

The following were major considerations in the design of the test tasks:

1. *Functions of the speaking test* (see Riggenbach 1998, Bygate 1987, O'Sullivan, Weir & Saville 2002)

   - Since the computer test should be equivalent as much as possible to the direct test, it is crucial that the speech functions found in the direct test are incorporated in the computer test

   - After viewing several clips of the direct test were (8 clips in total), using the observation checklist for speech functions (O'Sullivan, Weir & Saville 2002) to determine functions elicited by students during the group discussion task, it was decided that there were limited or no interactive functions, and no agenda management in the discussions

   - This implies that the discussion task in the direct test does not reflect 'real' interaction amongst candidates; the computer test can build in some of the functions that do not occur in a direct test and more functions can be elicited

2. *Co-construction of discourse in interaction* (see Lumley & Brown 1997, Brown & Hill 1998, Brown 2003, Luoma 2004)

   - In task B of the UiTM test, there is a danger that the discourse is co-constructed, as happens in most group discussions of this nature

   - The computer test can control this through similar/parallel input for all test takers. This is important and is one of the major problems to affect scoring validity, i.e. rating is affected when discourse is co-constructed by the speakers because the speech process and outcome are unpredictable. In addition, giving individual marks based on a group performance is a major problem for raters.

3. *Rating criteria* for the speaking test (see Eggins & Slade 1997, Abdul R. 2002, Hughes 2002/2003; Dimitrova-Galaczi 2004)

   - After comparing rating criteria from different tests: TSE, TEEP, IELTS, CPE, CELS, the new TOEFL standards for speaking was selected.

   Reasons for using the new scale for both tasks A & B:

   - Though staff suggested separate scales for the tasks, it is difficult & confusing for raters to switch between rating scales (especially with the criterion 'grammar')
   - Our test mimics the idea that a three-way conversation has 'chat' (short turns) and 'chunk' (longer turns) in the turns that the participants take (Eggins & Slade 1997); thus, whether for an individual presentation or an interactive task, the criteria for rating can be the same

- new TOEFL scale: though rather detailed, wordy & can be difficult to use (as it was developed for commercial use), is a good starting point for developing a more balanced scale for UiTM ; it is well-researched, is used for an established exam, has detailed descriptions of characteristics for each band, and is holistic in nature (practicality)

4. *The speaking test script* (see Appendix 3.11: Speaking test script task A/B)

- Considerations for test instructions (both tasks) and dialogue (task B)

  - Topic for part 1 (monologue) is thematically linked to part 2, more information is built in for test taker's schema so they have some ideas on what to say
  - Preparation phase/time built in for both tasks
  - Speech functions built into dialogue; more functions can be included in the computer test, including interactional functions which candidates are expected to recognise

    For example, in the direct test, candidates arrived at a decision with no clear interaction management but in the computer test, the interlocutor provides prompts by which the decision is to be made, such as:

    "But which one, a school or a university? OK, lets make a decision, which of the two should we go for, and what kind of things should they do there?"

    The function 'summarizing' and 'justifying' are also evident in the computer test:

"Ok...you spoke about places to visit, and we've just discussed things like food, a talk, and an educational visit. So, let's finally decide what to do. Can you summarize the ideas we've had and say what you think would be the best thing to do and why? "

C. On differences between the 'new' test compared to the 'original' test

- The differences between the two tests are described in the table below
- The main advantage of the computer test is it does not suffer from co-construction of discourse prevalent in the direct test

	Original test	New test
Turns	Long turns only with better students	same for all test takers
Length	Better students speak more times	same for all test takers
Functions covered	Limited/No interactive functions Limited/No agenda mgt	same for all test takers No agenda management

D. The Computerized speaking test: A Prototype

(see Appendix CD II attached)

The following table (Figure 5.1) illustrates different aspects of the test as they appear on the computer (see Appendix 3.12 for Computer test CV Specs for details of test context)

Figure 5.1 Features of the computerized speaking test

Test type	Objective	Format	Test input	Conduct
• Use of computer interface/ delivery; conducted in computer labs with stand alone machines, one per candidate • Candidates' responses are recorded in the installed MP3 recorder	To test a student's ability to speak in English while demonstrating informational and interactional functions, based on the tasks assigned to him/her in the test	• Like the UiTM test which had two tasks, a presentation followed by a group discussion, the computer test incorporates the two tasks: • Task A Candidates given instruction and information for the task, inclusive of preparation time, followed by presentation • Task B As in task A, candidates are given instruction and information for the task, inclusive of preparation time, they respond to prompts or questions posed by a speaker in a dialogue, as presentation	• Test instructions and information for tasks appear on the computer screen; • Candidates read and listen to them at the same time • In task B, a dialogue takes place and the script appears on the screen	•*Task A* 1. Candidates listen to and read instructions; they are able to think about the points they want to make during a 1 minute preparation time 2. Candidates record their presentations within 2 minutes speaking time •*Task B* 1. Candidates listen to and read instructions; they are able to think about the points they want to make during a 1 minute preparation time 2. Candidates respond to questions prompted by a speaker in an ongoing dialogue 3. . Candidates record a series of 4 short responses of 1 minute speaking time each

Note: The total time it takes for the whole test to complete is approximately 13 minutes, inclusive of preparation (2 minutes) and speaking time (6 minutes)

5.3.2 Test operations phase

The computer test went into operation in September 2004 when a pre-trial (monologue test) was conducted in Malaysia, followed by a pilot study in Feb-Mar 2005 (both tasks A: presentation, B: interaction). Following these trials and analyses of their outcomes, Main Study 2 was carried out in September 2005.

It is important to note at this point that each phase of the operations had a different focus or concentration. At the pre-trial, the focus was on the test itself, at the pilot study, the focus was on the questionnaires and processes, and at the Main Study phase, the focus was on all aspects of computerized testing, the test context validity, process or theory-based validity, and test reliability.

The details and description of these activities are described in Figure 5.2 below. (see boxes in yellow)

A. Pre-trial of the test

As was mentioned before, the main aim of the pre-trial stage was to establish the conditions and context for the computer test, i.e. to discover whether infrastructure, support and participants are available for the test to be administered; the focus at this stage is on the test itself. In order to achieve this, the researcher embarked on several measures:

- Focus group meetings with representatives from various academic and private institutions were held to discuss the possibilities of conducting the computer test. This involved issues related to the direct test, the proposal on a computer- test, and accessibility to participants, computer laboratories equipped with recording software such as Divace or Sound Forge, and other related support such as administrative and scheduling concerns.

- As per discussions above, the outcome was positive as all parties provided the researcher with access to students and computer facilities required. The computer trials were carried out at the respective centres and universities:
    - Five centres equipped with computer facilities, and approximately 100 participants were employed in the trials
    - At the end of each session, participants completed questionnaires relating to context and theory-based validity of the computer test.

B. Pilot study

The main aim of the pilot study was to check the feasibility of administering the test to candidates at the university (UiTM Malaysia), and to gather feedback from them on test context and processes involved in performing the test on the computer; the focus at this stage is on the test and processes.

In order to achieve this, the researcher embarked on the following procedures:

- With the help of lab assistants at the university, the test was reformatted to enable candidates to access it through the web and not have to manage it in its original format (see description of problems with the test in chapter 3, section II). This enabled each candidate to type in a URL and to have immediate access to the speaking test.
- Preparations were made for the pilot study:
    - A computer lab was reserved for the trials

- Schedules were arranged with lecturers who volunteered their students for the trials
- The computer test was revised by lab technicians: voice recorders were built in, instructions were revised, and test finally installed into the main server of the laboratory

- The test was conducted and data was gathered from participants as follows:
  - Total number of participants: 40, from four faculties and courses
  - Total number of hours for administration: 4+ hours over two days, inclusive of introduction and instruction/reminders to participants
  - At the end of each session, questionnaires (CV and TBV) were distributed for participants to complete
- A presentation on 'Developing criteria for rating a speaking test' was conducted at the university to gather data/ feedback from lecturers, including test setters and administrators involved in the speaking test.
  - Participants highlighted that problems usually faced by examiners were rarely discussed and addressed, including vague criteria and 'unfriendly' rating scale
  - Participants found the new TOEFL scale wordy but could be studied further for use at the centre
  - Participants found the computer test a bold challenge & were interested to examine it further

C. **Main Study 2**

Following the pre-trial and pilot study, preparations were made for the main study of the web-based speaking test. The aim of the main study was to gather evidence on the validity of the computer test; hence, the focus at this stage is on all aspects of the test including its context validity, theory-based validity and reliability. Data was gathered from the administration of the computer test to approximately sixty participants, context validity and theory-based validity questionnaires completed by the same participants, and rating of these tests using the new TOEFL scale by members of staff. After the same participants had completed their direct test in September, data was gathered from them in terms of the same sources of evidence for the computer test, i.e. context and theory-based validity.

- Preparations for the test included making reservations of the computer lab in advance of the sessions, test was revised by the lab technician & put on the web so that candidates have access to the test by entering the URL for the test, and arrangements were made with class lecturers for their students' participation in the test. The details are as follows:
    - Total number of participants: 61, from three faculties and courses
    - Total number of hours for administration: 6+ hours over three days, inclusive of introduction and instruction/reminders to participants

- At the end of each session, questionnaires (CV and TBV) were distributed & and completed by participants

- Test conduct:

1. Candidates were given a brief **introduction** on: purpose of test, format, time for preparation & presentation and **instructions** for the conduct of the test on: how to start, record, save audio files & end

2. Candidates were given reminders as follows:

   - Listen & read carefully before moving to next page

   - Use the preparation time to think of points and to take notes

   - Candidates were not allowed to discuss during the test

3. After the test, each participant was given the following **questionnaires** to complete:

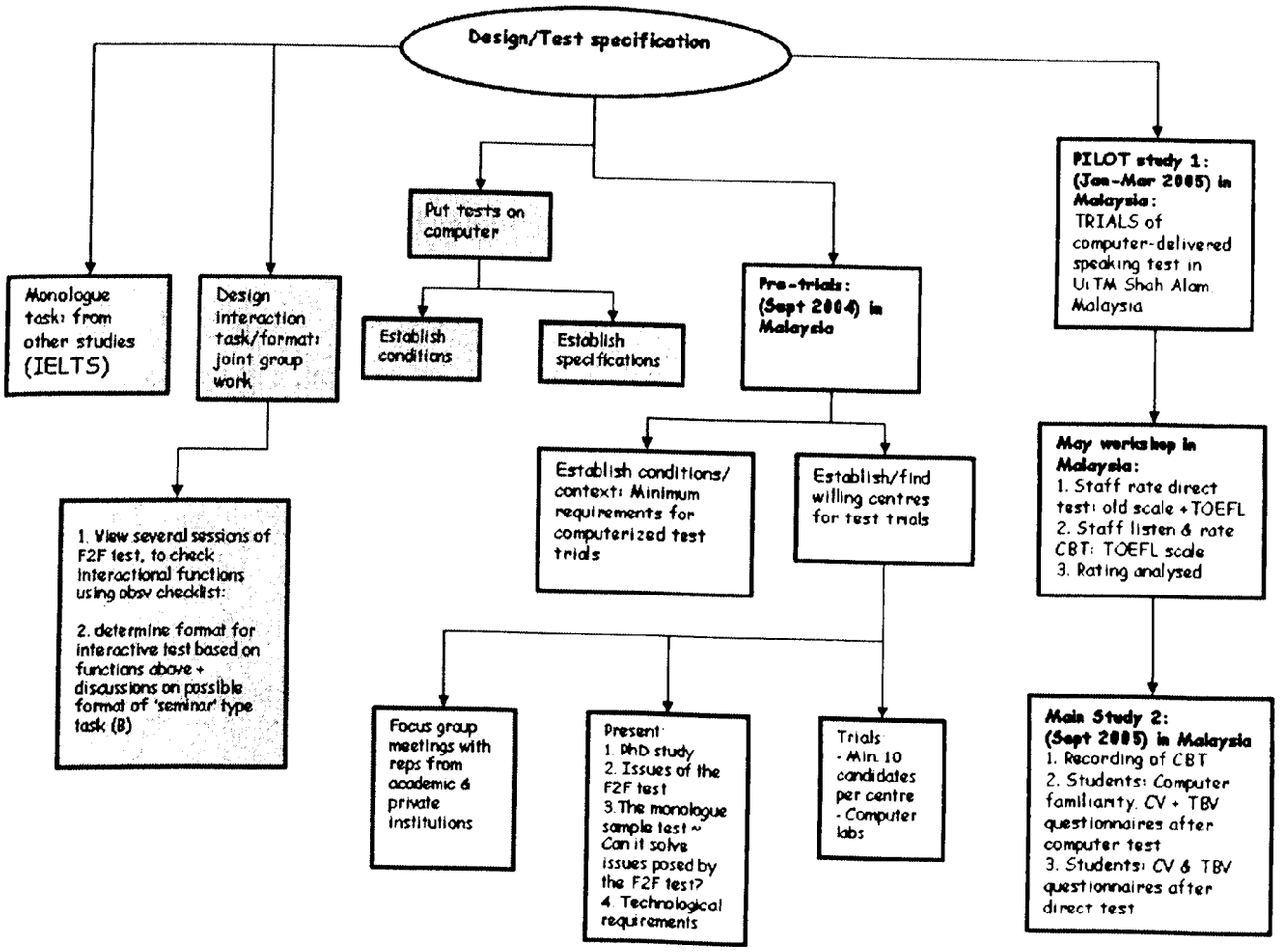
   a) Computer familiarity

   b) Context validity

   c) Theory-based validity

4. Further data were collected from the same candidates after they had completed the direct test (in September 2005) in terms of context and theory-based validity questionnaires

Figure 5.2 Model of computer test design and operations



5.4 FINDINGS FROM PILOT STUDY AND MAIN STUDY 2

In essence, we had determined from the pre-trial study (as detailed above) that support and infrastructure for delivering the computer test were available and accessible. At the pilot study stage we went on to conduct the completed first version of the computer test to the first sample of participants at the university, and data was collected using the context and theory-based questionnaires. These data and the findings were then examined so that the computer test may be refined and finalized for use in Main Study

2. Hence, the computer test did not only build upon data gathered from the first validation study of the direct test in its early development, but also data gathered from its own trials and pilot study for further refinement so that a final (prototype) form could be used in Main Study 2.

The findings from the pilot study and Main Study 2 are presented below. Like the direct test (chapter 4), the questionnaires used for the computer test validation were similar in terms of format and content, with minor changes relating to test delivery; all other elements remained the same as they are found in the framework for validating the speaking test (Weir 2005).

5.4.1 Data from Pilot study

The pilot study was conducted in Feb-Mar 2005 and at this point the test encompassed both tasks A: two minute monologue given the test task, and task B: four 'mini monologues' in response to questions or prompts from a dialogue in session, each one lasting one minute.

The following findings are based on participants' feedback on the test in terms of context and theory-based validity, both of which were gathered through questionnaires (see Appendix 3.9). A summary of the pilot study findings can be found in Appendix 3H in CD attached; however, only the **significant findings** are presented below.

In general, for the pilot test, the percentage of agreement on the validity elements for test setting and task demands, were high, some even higher than those found for the same elements in the direct test.

5.4.1.1 CONTEXT VALIDITY

TASK SETTING

Weighting: 40% disagreed on equal weights for both tasks
42% had no view/were uncertain
Note: This response was similar to the data found in Main Study 1; the candidates' preference here may be because the tasks are similar to the tasks for the direct test

Known rating criteria: Only 32% agreed this was made known
50% had no view/were uncertain
Note: At this stage of the computer test, the rating criteria were not built into the test

Security: Only 54% agreed that students were not able to discuss test questions
Note: At this stage of the study, though each participant sat at individual computer stations, all thirty stations used were in one laboratory; it was clear that candidates were able to talk to each other, even though they were instructed not to.

Computer-delivered test: Only 24% preferred
40% did not prefer
34% were uncertain
Note: As noted in Main Study 1 where the figures for this item were fairly low, this was the first time participants would have been exposed to a computer-delivered speaking test, in spite of data on computer familiarity (see findings MS2 below) which indicated high levels of knowledge and familiarity on computers and computer usage.

TASK DEMANDS

Topic familiarity: Only 50% agreed that the topic was familiar

Interlocutor variables:
Elements with low percentage of agreement, or inconsistency are:

Examiner help: Only 54% agreed that examiner in the test was helpful
Gender: In task A, 50% preferred the same gender/ 50% did not
In task B, 50% had no view, 19% did not prefer the same gender, 29% did

Note: Overall, candidates did not have a problem with examiner/interlocutor speech rate, accent, acquaintanceship; the figures above reflected data found in Main Study 1 in which examiner/interlocutor gender response was mixed, and so was examiner help.
In MS1, this was due to inconsistencies in test administration a cross test centres; in the computer test, this does not happen.
Perhaps the term 'examiner help' in the questionnaire for the computer test was vague and needs to clarification.

5.4.1.2 THEORY-BASED VALIDITY

EXECUTIVE PROCESS

Task A:

In terms of **before starting the test**, percentages were low for:
'thought of how to satisfy the examiner' and **'thought of structures needed'**:
58% agreement on both items, i.e. only 58% agreed they did them
Note: It is interesting to note a similarity in feedback for both items to Main Study 1 where the percentages were low too. However, for 'thought of how to satisfy examiner' the percentage was higher in the direct test, probably because the examiners were present at the test, but in the computer test, they were not.

In terms of **during planning time**, percentages were low for:
'thought of structures needed': 42%
'thought only in my own language': 42%
'thought in both English & own language': 55%
'able to put ideas/content in good order': 25% disagreed; 58% uncertain
Note: When candidates are confronted with these questions in relation to the new computer test, it may be difficult for them to express their thought processes and strategies. Hence, there were some uncertainties/no view response in this section. The highest percentage of concern here is in the last item, i.e. more than 80% were not able to or were uncertain about putting ideas in good order.

In terms of **while speaking**, percentages were low for:
Checked grammatical accuracy: 55% agreed they did this
organization of presentation: 45% agreed they did this
Adjusted grammatical accuracy: 55% agreed they did this
organization of presentation: 47% agreed they did this
Note: In addition to the fact that the computer test is a new test for all the participants, candidates often underestimate their abilities; however, data for Main Study 2 (section later in this chapter) seem to show more positive results.

Task B:

In terms of **before starting the test**, percentages were low for:
'thought of how to satisfy the examiner': 50% agreed they did this
'thought of structures needed' : 58% agreed they did this
'information in A helped prepare for B': 58% agreed
Note: Similar to data above for task A, candidates are uncertain about their thoughts and abilities before starting the test. It is encouraging that more than 50% found the information in the test helped think about the task at hand.

In terms of **during planning time**, percentages were low for:
'thought of structures needed': 47%
'thought only in my own language' : 42%
'able to put ideas/content in good order': 61% disagreed
Note: When candidates are confronted with these questions in relation to the new computer test, it may be difficult for them to express their thought processes and strategies. Hence, there were some uncertainties/no view response in this section; though the percentage is higher here than in task A, for 'able to put ideas/content in good order'.

In terms of **while speaking**, percentages were low only for the item:
Checked grammatical accuracy: 47% agreed they did this
Adjusted grammatical accuracy: 55% agreed they did this

Note: In this part of processing candidates seemed more confident about other elements such as checking word use, points they wanted to make, points the other speakers make, and adjusting their response based on what other speakers said; all this even in the computer test.

EXECUTIVE RESOURCE

The only problem element for the participants here is in internal content knowledge:
Information for the task was familiar from previous readings and experience
42% agreed; 13% disagreed; 45% were uncertain
Note: It appeared that candidates were confident about topic knowledge, functional and sociolinguistic knowledge of the tasks in the computer test. These data are positive indicators of the clarity and appropriacy of test instructions and information provided for the test tasks, including the dialogue in task B. For example, candidates indicated that they were able to 'connect what they said to what's been said' in the dialogue (82% agreement). In addition, percentages for executive resource elements in this survey were higher than those for the direct test, such as 95% (task A) & 92% (task B) for 'information in the instructions were necessary to complete the task'.

CONCLUSION

In general, participants' response to the context validity of the computer test showed an improvement over data from Main Study 1of the direct test in areas such as clarity of purpose and order of items, time constraint, test administration and conditions, interlocutor variables, channel of communication, and linguistic range of information.

For theory-based validity, though the elements that candidates seem to have problems with are the same as those found in the direct test such as thinking of or checking the structures needed in their speech, and organization of speech in good order, they appeared to have less difficulty with executive process for task B, and executive resources such as topic knowledge, functional and sociolinguistic knowledge. Therefore, in this first pilot study of the computer test with both tasks A and B, candidates seemed to be able to complete the tasks, in spite of having doubts about their own abilities and dealing with a new method of testing their speaking abilities. The

response regarding the test so far, in terms of context and theory-based validity indicates emerging positive aspects of the computer test.

In Main Study 2, all these aspects come to focus, i.e. the test, the questionnaire used for feedback, and test scores are taken into consideration when validating the computer test.

5.4.2 Data from Main Study 2

(see Appendix 3L in CD I attached for SPSS outputs; Appendix 3H for a summary of the findings)

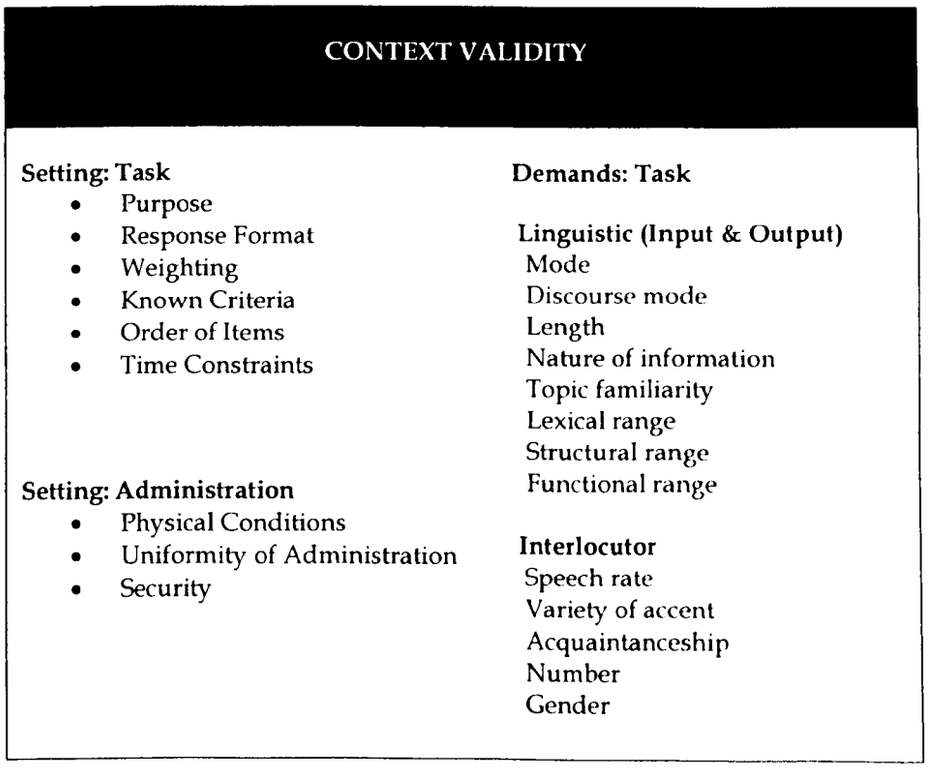
Main Study 2 began in May 2005 when a workshop on testing speaking was conducted with members of staff (lecturers, examiners, administrators, experts) at the Language Centre. Data from student participants were gathered in July - September of the same year. In both occasions, the participants were given questionnaires: students (context, theory-based validity), staff (context, scoring validity), some students wrote down comments as regards the computer test, while staff members provided feedback on the computer test in terms of its strengths and weaknesses, and rating.

The findings, based on all participants' feedback in terms of context, theory-based and scoring validity of the framework for validating a speaking test (Weir 2005) are presented below in three parts. In each part we present again the **table of validity elements**; the **definitions for each element are found in the findings of the direct test in chapter 4 and so will not be repeated**. As in data from Main Study 1 (chapter 4), data

for the computer test is reported in terms of each validity element of the framework. The final data includes an analysis of test scores of candidates who participated in the computer test and the direct test, in terms of basic correlations and test reliability.

It is important to note that the computer test was developed based on data from Main Study 1, and that it parallels the direct speaking test in every aspect; hence, the validity elements by which the findings of Main Study 2 are reported below reflect the same elements of the direct test. More importantly, as stated earlier in this chapter and previous chapters, the computer test parallels the direct test in all elements of context validity. (see Figure 5.3 below). Hence, the specifications for the test are similar in many ways, such as test purpose, order of items, time constraint, topic familiarity and other task demands (Appendix 3.12).

Figure 5.3 Aspects of Context Validity for Speaking (from Weir, 2004)



5.4.3 CONTEXT VALIDITY

5.4.3 a) Task Setting

Purpose The purpose of the test is to test a student's ability to speak in English while demonstrating informational and interactional functions, based on the tasks assigned to him/her in the test.

Table 5.1a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	95% agreement that purpose is clear
Questionnaire	Lecturer/Examiner (N=41)	100% agreement that purpose is clear
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview (group)	Staff (N=30)	Equivalence of input for all candidates in terms of instructions is good(speed, accent, linguistic & functional features are the same for all); makes purpose clear

Commentary

All participants are in agreement that the purpose of the test was clear. At the group interview/discussion, staff members indicated that equivalence of input is an advantage of the computer test in that everyone gets the same input. This is turn addresses the issue of co-construction of discourse which is a major concern of the interactive task in a direct speaking test (Brown 2003/2004, Fulcher 2003, Luoma 2004).

Overall, the purpose of the test is clear to all candidates, in spite of the fact that for most participants, this was the first time they had viewed the test in its entirety.

Table 5.1b

Document Analysis	
Document	*Findings*
CB test description/design	Purpose of task A & B stated
CB test script	Clear in the instructions of the tasks
Rating criteria (new TOEFL)	N/A

Commentary

The document CB test description/design describes the main purpose of the test which is to test a student's ability to speak in English while demonstrating informational and interactional functions, based on the tasks assigned to him/her in the test. This information is translated in the test script/instructions and information that students read and listen to during the test. While the rating criteria (new TOEFL) was not developed for this speaking test, it was selected mainly because it has three criteria which are similar to the old scale and as such reflect the model of linguistic knowledge expected of the successful candidate, and although its association with the computer test is also not direct, it is an established scale, and the computer test was developed based on the literature and empirical data.

Table 5.1c

Observation	*Findings*
Researcher	Students behaviour during the test sessions indicated they were able to complete the test without major problems understanding the tasks

Commentary

Candidates indicated they understood the tasks fairly well as no one asked for assistance regarding test instructions and tasks; most of the assistance needed was for technical reasons such as recording and saving of files.

Response format for the computer test consists of the completion of two tasks: a long turn monologue (similar to task A in direct test), and shorter turns where they respond

to prompts or questions posed to them through a dialogue (similar to task B in the direct test).

Table 5.2a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	For task A: 92% agreed it test communication in academic context, 87% in social context For task B: 92% agreed it test communication in academic context, 85% in social context
Questionnaire	Lecturer/Examiner (N= 41)	For task A: 63% agreed it test communication in academic context, 54% in social context For task B: 44% agreed it test communication in academic context, 46% in social context
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview (group)	Staff (N=30)	No comment on response format

Commentary

Because the original speaking test specifications had no clear and detailed information regarding all elements relating to test context, participants may still be confused with this aspect of the test. However, it is clear in the computer test that candidates perform one long turn monologue, followed by four shorter turns. While students' agreement on this is high, staff figures are lower, especially for task B. It is unfortunate that there is still confusion amongst staff on this point, though it is understandable as they had been involved in the direct speaking test for many years. The relationship between response format (especially task B) for the direct test and this test may still be unclear to them. In general, the response format for the test was adequate for the students but not so for staff members.

Table 5.2b

Document Analysis	
Document	*Findings*
CB test description/design	Evident in the document
CB test script	Response format is clear
Rating criteria (new TOEFL)	N/A

Commentary

The response format is clearly described and expressed in the documents above, except for the TOEFL scale, which is not directly related to the test after all.

Table 5.2c

Observation	*Findings*
Researcher	Students had no problems relating to response format; they seemed to be able to perform the long turn and shorter turn monologues within the time limits

Commentary

Candidates seemed to be able to handle the tasks fairly well, performing the long turn presentation followed by shorter turn's responses. It was encouraging to see them attend to the tasks without questions relating to response format.

Weighting For the computer test, the tasks have equal weighting, as they were originally stipulated in the direct test; no changes were made here.

Table 5.3a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	Only 51% agreement on this
Questionnaire	Lecturer/Examiner (N= 41)	73% agreed that the tasks should be equally weighted
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview (group)	Staff (N=30)	Not sure, need to be thought through and discussed again

Commentary

Staff agreement on this element is probably justifiable seeing that they indicated in table 5.2a above that there is no real difference between the tasks, and in the discussion they indicated that this needs to be considered and thought through, now that 'interactive' property in task B has changed. It was clear for this element of the direct test (ch 4, section 4.3.1, table 4.3a) that the pattern is similar, i.e. students' agreement was lower than staff, though the percentages here are higher.

In general, weighting may be a problem for the participants at this point of the test as the relationship between tasks may not be immediately clear to them.

Table 5.3b

Document Analysis	
Document	*Findings*
CB test description/design	Not evident, though implication is that there's no change from the original test
CB test script	Not stated here
Rating criteria (new TOEFL)	N/A

Table 5.3c

Observation	*Findings*
Researcher	Nothing was raised at the test sessions regarding weighting.

Commentary

During the observations, no questions were raised about weighting of tasks. This however did not have an effect on the students who proceeded with the test, with minimal questions relating to test context.

Known rating criteria There are three criteria for rating this test and they were made explicit in the test: language use, delivery, topic development. Each one was explained to the candidates as they read and listened to instructions for the test.

Table 5.4a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	85% agreed the criteria were clear in the test
Questionnaire	Lecturer/Examiner (N= 41)	95% agreed the criteria were clear in the test
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview (group)	Staff (N=30)	No comments here

Commentary

The percentage of agreement for this item is high amongst all participants as the criteria were clearly listed and explained in the test, even though there was a difference from the criteria in the old scale (e.g. 'communicative ability' in old scale but 'delivery' in the new scale). Hence, it is obvious to all parties concerned that the criteria for rating have to be made known, especially to the test takers, examiners and raters, and as clearly as possible.

This element is evident in the test and is not a problem here.

Table 5.4b

Document Analysis	
Document	***Findings***
CB test description/design	Criteria for rating available
CB test script	Criteria for rating made explicit in the test instructions
Rating criteria (new TOEFL)	3 criteria: Language use, Delivery, Topic development Each one had detailed descriptions and placed on a band of 1(lowest) to 4 (highest)

Commentary

All the documents contained descriptions of the criteria for rating the tasks; the test description listed the criteria for rating, the test script had explicit descriptions for the test taker, and the rating criteria had detailed descriptions and how they are placed on a scale. Hence, necessary information relating to rating criteria must be made explicit to all parties concerned directly or indirectly to the test

Table 5.4c

Observation	*Findings*
Researcher	Students did not raise questions or made comments relating to the criteria for rating

Commentary

Again, candidates seemed to be able to participate in the test without problems relating to known criteria for rating.

Order of items While the order of the tasks in the original test (task A followed by B) was related to a thematic link between them, this was maintained in the computer test, even though 'interaction' in task B is different from the way it takes place in the direct test.

Table 5.5a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	92% agreed the order was appropriate
Questionnaire	Lecturer/Examiner (N= 41)	95% agreed the order was appropriate
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview (group)	Staff (N=30)	No issue here as it is the same in the direct test

Commentary

All participants agreed that the order of the tasks was appropriate, i.e. task A followed by task B. In the original test, the order was pre-determined by the thematic link in that the points raised in task A influenced the outcome of task B. In the computer test, this was maintained, and even if it were not apparent to the student because it was 'interactive' only in that they had to respond to questions prompted by the interlocutor on the screen, the order was not affected.

Hence, this aspect of the test is not a problem for everyone concerned.

Table 5.5b

Document Analysis	
Document	***Findings***
CB test description/design	Order of tasks made clear
CB test script	Evident in the test: candidates need to perform in the order task A, followed by task B
Rating criteria (new TOEFL)	N/A

Commentary

Information regarding order of items is made clear in the specifications and script for the test; this is parallel to the original test in which the order was pre-determined. In the computer test, it is also not possible for candidates to do the test in a different order, especially since order affects processing and they needed to complete task A before they could proceed to task B and complete the task.

Table 5.5c

Observation	*Findings*
Researcher	Students were able to perform the tasks according to the order they appeared in the test

Commentary

Students did not seem to be affected by the order the tasks were presented in the test, i.e. how they appeared on the screen. This aspect was controlled for everyone and this is also an advantage of the computer test in which the tasks are built into the system such that all candidates perform the test in a specific order.

Time constraint: For the computer test, preparation time is 1 minute for each task, and presentation time is 2 minutes for task A, 1 minute per response for each of the four times the candidate speaks in task B.

Table 5.6a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	On **presentation time**: task A: 74% agreed the time was sufficient task B: 71% agreed the time was sufficient
Questionnaire	Lecturer/Examiner (N=41)	On **presentation time**: task A: 88% agreed the time was sufficient task B: 80% agreed the time was sufficient
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 13)	Students commented on insufficient time for **preparation**, i.e. 1 minute for each task is not enough, especially since it is a new method of testing
Interview (group)	Staff (N=30)	Equal time for everyone to prepare and respond; control here is ensured

Commentary

All participants seemed to agree on the time for presentation, though students commented on the time for preparation, which they found too short, and especially since this is the first time they are doing the test via the computer, which made them more anxious. Many examiners and lecturers were satisfied that the computer is able to control the amount of time for preparation and presentation for each candidate, so there is no variance in terms of test administration.

Table 5.6b

Document Analysis	
Document	***Findings***
CB test description/design	Both time for preparation and presentation stipulated in the specifications
CB test script	Both time for preparation and presentation made clear to candidates as they read and listen to instructions
Rating criteria (new TOEFL)	N/A

Commentary

The time constraint for both preparation and presentation of the tasks are made clear to all parties concerned in the test specifications and the test script itself.

Table 5.6c

Observation	***Findings***
Researcher	Students were able to complete the test according to the time stipulated

Commentary

In general, candidates were able to fulfil the tasks without major problems with the time given to them to complete the test. As mentioned above in table 5.1c, much of the assistance they needed was mainly for technical help, such as in recording their presentations and saving them in specific folders.

5.4.3 b) Test administration

The computer test was conducted in computer laboratories equipped with modern facilities. Each computer lab has five-computer group stations (five stations in a lab), each computer has direct access to the internet, communication between candidates is possible, facilitator communication from the main controls is possible with individual candidate and group, and other facilities such as high quality audio controls and wide screen projection of video presentations. The labs are fully air-conditioned, have good acoustics, are well lit at all times, and qualified lab technicians are available.

Physical conditions: Candidates are asked if physical conditions such as lighting, noise level, room temperature, seating arrangement and conditions for disabled were satisfactory and appropriate for a test of this nature

Table 5.7a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	Lighting: 90% agreement Noise level: 82% agreement Room temperature: 74% agreement Seating arrangement: 85% agreement Conditions for disabled: 59% agreement
Questionnaire	Lecturer/Examiner (N= 41)	N/A as staff members were not present during the test administrations
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	

Commentary

Overall, candidates felt that the conditions in the labs during the test were appropriate. Candidates commented after the test sessions that they enjoyed the test, and attendance was not a problem each time the test took place in the labs. In terms of conditions for disabled, students were uncertain if they were available; hence, this is one aspect that needs to be improved such as conditions for dyslexics or students confined to a wheelchair.

Table 5.7b

Document Analysis	
Document	*Findings*
CB test description/design	Test conditions and environment described in specifications
CB test script	N/A here
Rating criteria (new TOEFL)	N/A here

Commentary

Conditions for conducting the test in terms of venue, facilities and other conditions are spelled out in the test specifications but not in the test script and rating criteria. Overall, specific conditions have been listed and should be met.

Table 5.7c

Observation	*Findings*
Researcher	Students appeared to be able to take part in the test sessions comfortably, though at times there were too many of them seated in one lab

Commentary

Most students appeared at ease during the test, each one busy fulfilling the tasks at his/her station. There were sessions when there were too many candidates because of timetabling, and they were seated too close to each other.

Overall, physical conditions for the test were appropriate and satisfactory.

Security, especially for web-based tests, is a big concern because technology has enabled us to infringe on confidentiality and security of any materials published on the web. In the new computer test, candidates are asked if they were able to discuss test questions; unfortunately, during the pilot and Main Study trials, there were opportunities because of physical arrangements in the lab. Eventually, this could be controlled and other measures for confidentiality and dissemination of test questions and recordings need to be addressed. In the case of UiTM, this is a major concern because security was an issue with the direct test. (see finding in chapter 4)

Table 5.8a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	Only 57% agreement
Questionnaire	Lecturer/Examiner (N= 41)	Not available here
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview (group)	Staff (N=30)	This could be a serious issue with the computer test

Commentary

In general, like any other test, the concern with test security is a valid one because the computer test is new and this is the first time it is conducted at the university. Students realized that they were able to talk to each other even though they were seated apart from each other in the labs. Staff members were understandably concerned about how test security is maintained with the computer test.

Test security is one aspect of the test that needs a lot of attention, even though security measures can be easily put in place with the technology available. Eventually, candidates can only access the test with a user name and password, the completed test is saved in a local server before it is sent out for rating via internet files, and so on.

Table 5.8b

Document Analysis	
Document	*Findings*
CB test description/design	Information not available yet
CB test script	
Rating criteria (new TOEFL)	

Commentary

Information on test security is not available in the documents even though it was discussed with members of staff. Further discussions with administrators and technical specialists will enable us to develop a system for ensuring test security, and this Information would be include din the test specifications.

Table 5.8c

Observation	*Findings*
Researcher	The seating arrangements in the lab were organized ahead of the test; candidates were well spaced out, even though at times they sat in closer proximities when the number of candidates present was bigger

Commentary

Test security is a key problem in web-based tests, especially for speaking. In spite of the fact that literature on security of such tests is scarce, it is a major concern for this test. Observations during the test sessions showed that candidates could talk to each other even during the test, especially if they were seated close to each other.
Eventually, considerations would be taken regarding number of candidates in a lab, technical specialist for organization and dissemination of student files to raters, and security of test scores.

5.4.4 OVERALL COMMENT ON TASK SETTINGS

Thus far, the context validity in terms of task setting for the computer test looks promising. Overall, all participants concerned, and test candidates specifically, seemed

clear about test purpose, response format, known criteria, order of items, time constraint, and physical conditions of the test. The items weighting and security are major concerns which need to be addressed further following the outcome of the study. However, weighting was an issue raised in Main Study 1 of the direct test in which the justification for the tasks having equal weighting was not clear. Because the computer test tasks are parallel to the direct test, the weighting for each task remained the same. While data for both tests indicated that participants preferred each task to have different weighting, the concern is in the design stage where this element needs to be made clear to test developers, examiners, lecturers and test takers themselves; only then are we able to decide if the weighting is appropriate.

In terms of security, I feel that technology is available to safeguard the confidentiality and security of test documentation, scores and so on, as literature on computer-based or aided testing is growing. (For details on advantages/disadvantages and major concerns/issues of CBT/CBA see Kenyon et al 1999-present; Fulcher 2000, 2003; Thelwall 2000; Russell 1999, Russell & Haney 2000; Chapelle 2001, 2003; for comparisons between computerized and other conventional tests see McDonald 2002; Neuman 1998; Russell & Haney 1997; Mead & Drasgow 1993; articles in Language Learning & Technology: http://llt.msu.edu, and Henning 1991; Cumming et al 2004; Stricker 2004 on validating CBT/CATs).

One of the major advantages of the computer-delivered speaking test is in standardization. This relates to the problems raised in the direct test such as equivalent forms or input, co-construction of discourse in interaction and test reliability caused by

inconsistencies in test administration and test conditions. With the test delivered through the computer, the following elements are standardized for all test takers: time constraint (preparation and presentation), response format, order of items, channel of communication, topic/content knowledge required, nature of information and test administration. All candidates are given the test in similar settings, administered in the same manner, and attend to task demands that are the same for everyone. The issue of unfairness due to variance in topic or task difficulty does not arise in the computer test. In task B, because candidates respond to specific prompts and are asked specific opinions and ideas, the problem of co-construction of discourse between speakers do not occur; students are able to demonstrate their ability and fulfil the task functions without being influenced (positively or negatively) by other speakers. Indeed, through this mode of delivery, candidates can begin to feel less anxious and stressed due to variations in interlocutor variables and inconsistent test administration across campuses; the examiner/interlocutor are not present, and candidates may even have a choice of variables such as interlocutor gender and accent in the future.

5.4.3 c) Task demands:

Channel of communication: Written and spoken instructions

Table 5.9a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	97% agreed that having both instructions was helpful
Questionnaire	Lecturer/Examiner (N= 41)	95% agreed that having both instructions was helpful for candidates 85% agreed that information provided in instructions was sufficient for candidates
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview (group)	Staff (N=30)	Information is sufficient and complete

Commentary

All participants seemed satisfied with the channel of communication for the test in that students were given both written and spoken instructions, and more importantly that information provided in the instructions was sufficient for them to fulfil the tasks.

In general, the channel of communication in terms of the way the instructions were delivered and the amount of information provided are an advantage of the test.

Table 5.9b

Document Analysis	
Document	*Findings*
CB test description/design	How information is delivered and amount of information in instructions and test tasks are described in specifications
CB test script	Evident here
Rating criteria (new TOEFL)	N/A

Commentary

The test specifications contain descriptions of the information provided in test instructions and test tasks, and the channels of delivery, i.e. candidates read and listen to the information provided at the same time.

Table 5.9c

Observation	*Findings*
Researcher	Students appeared at ease about seeing the instructions on the screen while listening to them at the same time; few needed assistance with instructions, for e.g. about proceeding to the following page after the time for speaking is over

Commentary

One of the reasons for having both channels of delivery (written and spoken) at the same time is to prevent misunderstandings of instructions. It was clear from the test sessions that some candidates needed assistance as they were uncertain about procedure, rather than test instructions or test task, even though the procedural aspect of the test was incorporated in the instructions, for e.g. ... after BEEP sound: "Please stop talking now. You may proceed to the next page", and so on. However, this aspect of the test will need to be revised to make instructions clearer for candidates in the future. And even though data here suggest that candidates were not disadvantaged, the effect of multiple input on performance in speaking tasks is likely an increase in code complexity and this can make processing the task more difficult (Skehan 1998, Weir 2005).

The computer-delivered test Candidates were asked here if they preferred the computer test over the direct test; for the candidates in UiTM this would be their first attempt at doing the speaking test using the computer.

Table 5.10a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	No: 21% Yes: 49% 30% uncertain
Interview	Student (N = 13)	• Because it is a new system, it takes getting used to • Great that you could play back and listen to your own speech; good for practise, hence good way of building confidence, especially for shy students
Interview (group)	Staff (N=30)	Could address problem of standardization in administration of test across faculties & branch campuses

Commentary

Though the percentage of agreement on preference for the computer test is below 50% amongst the students, two points are clear. First, the percentage here shows an increase from the pilot data (see section 5.4.1 above), and secondly, in spite of 30% uncertainties in the response, candidates pointed out some positive aspects of the computer test. They were conscious that this was a new system which required some "getting used to" (candidates words), and that with sufficient practice it could even build confidence of the shy students who do not usually volunteer to speak in face-to-face interaction. In fact, data on rating the computer test showed instances of improvement in a candidate's

speech as they progress from task A to various stages of task B, something that happens in direct test setting as well. (see Appendix 3.10; findings 'scoring validity' below)

Table 5.10b

Document Analysis	
Document	***Findings***
CB test description/design	Information in instructions and test tasks and how this is presented on the computer are described
CB test script	Evident here
Rating criteria (new TOEFL)	Information found here regarding the relationship between the scale and how it was developed for the computer test

Commentary

The documents contain detailed information relating to the test instructions, test tasks and rating scale for the computer-based speaking test, i.e. the interface of information, how candidates attend to/perform the tasks using the computer, and how the performances are rated.

Table 5.10c

Observation	***Findings***
Researcher	Students were able to attend to the tasks with minimal problems; the assistance needed were mainly for technical problems such as recording and saving files

Commentary

More than six hours of test recordings in the computer labs were observed and in general, candidates were able to attend to the tasks with minimal help from the technical support assistants and the researcher. As mentioned before in tables 5.1 - 5.6

and table 5.9 above, candidates seemed capable of managing the test using the computer, in spite of their first experience participating in such a test.

Thus, candidates are reasonably confident about the test and seemed to face minimal problems in terms of the tasks and interacting with the computer.

Nature of information in tasks A & B

Table 5.11a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	Not available
Questionnaire	Lecturer/Examiner (N= 41)	34% agreed that students argue for and against an idea; 66% disagreed
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	

Commentary

The item in the questionnaire for this element was presented to staff participants, as they were directly responsible for the nature of information in the direct test, and the computer test tasks parallel the tasks in the direct test. In our finding for the direct test, administrators and other staff members agreed that the tasks were mainly experience-based rather than factual or abstract in nature, and not argumentative even though in the discussion task, this had to be demonstrated in order to support for or against ideas. In the computer test, task A is not argumentative in nature, and task B is argumentative to the extent that candidates have to justify, elaborate, and make a decision. In fact, it is probably clearer to administrators, examiners and lecturers that the nature of

information in the tasks is not argumentative, that this is different from the functions demanded of the students from the task itself/input, from the computer test. Hence, nature of information of the tasks is clearer in the computer test than it was in the direct test.

Table 5.11b

Document Analysis	
Document	*Findings*
CB test description/design	Nature of information described in the test specs
CB test script	Evident here
Rating criteria (new TOEFL)	N/A

Commentary

The nature of information of the test tasks is described in the test specifications and is evident in the script for the test.

Table 5.11c

Observation	*Findings*
Researcher	Like the findings for the direct test, it is hard to say if candidates were affected by the nature of information, until further evidence is gathered. However, at the test sessions, students did not raise questions regarding the information for the tasks.

Commentary

Again, it was difficult to say if candidates were definitely affected by the nature of information until further evidence is gathered. However, it is important to ask the

question if the type of information is appropriate for the target situation requirements of the students being tested (Weir 2005). Thus far, the candidates appeared to be able to perform the tasks without major problems in understanding the information presented to them, even if the lecturers and examiners were not certain of the type of information.

Topic familiarity

Table 5.12a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	71% agreed that topic is familiar
Questionnaire	Lecturer/Examiner (N= 41)	93% agreed that topic is familiar for candidates
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 13)	No problem with the topic; some have seen it in the test before
Interview (group)	Staff (N=30)	Some were familiar with the topic from a past paper

Commentary

Again, it is interesting to note that the percentages show an increase from the data for this element in the direct test, both for students (previously 61% agreement) and staff (previously 55% agreement only). The topic was selected from past papers of the direct test and hence the familiarity element, especially for examiners and lecturers. However, it is an improvement in terms of the findings, and the high agreement amongst staff

shows that they find the topic appropriate for the purpose of the tasks and for the students.

Thus, the topic appears not to be a problem for the computer test.

Table 5.12b

Document Analysis	
Document	***Findings***
CB test description/design	Topical range described in test specifications
CB test script	Evident here
Rating criteria (new TOEFL)	N/A

Commentary

The range of topics or type of content knowledge required of the students is found in the test specifications and evident in the test script itself. However, there is no information related to topic/content knowledge mentioned in the rating scale.

Table 5.12c

Observation	***Findings***
Researcher	Students seemed to have no problems with the topics for the tasks

Commentary

As with the other elements above such as response format, order of items, topic/content knowledge, and nature of information, all candidates appeared to have minimal problems while performing the tasks. As mentioned above too, most of the assistance needed was relating to technical matters.

Linguistic variables of text (in main text, test instructions, or task). In the computer test, this includes both the written (as presented on the screen) and spoken text (candidates listen to the text read out). These include lexical, structural and functional range of the text.

Table 5.13a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	72% agreed that **lexical** range is appropriate 92% agreed that **structural** range is appropriate 89% agreed that **functional** range is clear
Questionnaire	Lecturer/Examiner (N= 41)	95% agreed that **lexical** range is appropriate 98% agreed that **structural** range is appropriate 83% agreed that **functional** range is clear
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	

Commentary

The percentages for linguistic variables here are higher than those reported for the same elements in the direct test, and data from the pilot study; in both cases, the percentages were in the 70s. Hence, it is encouraging to see that most participants, students and staff, found the demands of the test in terms of lexical, structural and functional range to be appropriate and clear for the candidates taking the test. This is an advantage of the computer test; the linguistic variables are controlled and kept the same for all test takers. This is of importance in relation to the direct test where the linguistic range of the test

tasks were at times not appropriate, and was constantly changing. This is a problem especially for equivalence of test forms of six parallel sets that are produced each year.

Table 5.13b

Document Analysis	
Document	*Findings*
CB test description/design	Linguistic range of test instructions and tasks are described in the specifications
CB test script	Evident in test script (written) and spoken text
Rating criteria (new TOEFL)	N/A

Commentary

The test specifications and test script contain descriptions of the linguistic range/input of the written and spoken text as the students recognize them during the test; these are linguistic range of the test text, both written and spoken. As discussed in chapter 4 on findings for the direct test, the linguistic range of the test instructions and other information for the tasks must be appropriate for the level of the candidates.

Table 5.13c

Observation	*Findings*
Researcher	Students appeared comfortable with the input they received for the test (written & spoken text); no questions were raised in any of the linguistic variables

Commentary

As reported before, students did not seem affected by the input they received during the test in terms of written text on the screen and spoken text that they hear. They seemed capable of performing the tasks without help in comprehension of the instructions

and/or the test tasks. The advantage of the computer test is that the appropriacy and range of linguistic variables can be maintained each year while the topics and content change.

Hence, the linguistic variables/range of the test instructions and test tasks were not a problem for these candidates and were appropriate for their levels.

Interlocutor variables for the computer test refer to spoken instructions (both tasks A & B) and the speakers found in the dialogue in task B. The elements are similar to the ones found for the direct test.

Interlocutor variables for task A:

Examiner help, speech rate, accent; acquaintanceship ; gender ; examiner presence

Table 5.14a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	67% agreed **examiner help** was necessary 87% agreed examiner **speech rate** is appropriate 62% agreed examiner **accent** not difficult 62% agreed **acquaintanceship** is important 80% agreed **gender not a problem** 67% disagreed they preferred **same gender**
Questionnaire	Lecturer/Examiner (N= 29)	10% agreed **examiner help** was necessary for students; 61% disagreed 49% agreed examiner **speech rate** is appropriate; 19% disagreed 61% agreed examiner **accent** not difficult; 31% agreed **acquaintanceship** is important; 34% disagreed 10% agreed **gender is a problem** for students ; 20% agreed it is not; 34% were uncertain 51% disagreed students preferred **same gender**; 12% agreed Yes:29% Uncertain: 34% to **examiner not present preferred**
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview (group)	Staff (N=30)	• Tasks seem less stressful for candidates, especially with absence of examiner & other speakers • Advantage: no need for interlocutor training

Commentary

It is noticed that there were missing data from staff members on interlocutor variables (see Appendix 3H in CDI for MS2 Findings). Either this was due to questionnaires that were not returned or participants not present at the workshop sessions. However, for most elements, we received 60% of the response. For members of staff, the only similarity in response to the students is for examiner accent where 61% agreed this is not a problem for candidates. Response to other elements: examiner help, speech rate,

acquaintanceship, gender, were different from the students', mostly much lower agreement.

Overall, students felt that interlocutor variables for the computer test were appropriate and did not affect their performance in a negative way; staff members however, felt that these were concerns for the students. These findings are probably not surprising for several reasons. Firstly, the students participated in the test themselves. They experienced the effect of these variables first hand and were able to express this in the questionnaire. While few lecturers were present at the trials, others viewed the test at the workshop sessions and participated in the rating exercise. However, their uncertainty about the computer test is not surprising as this was the first time the speaking test was presented on the computer; most lecturers know of the direct method of testing speaking only. In fact, some commented that students sounded less stressed, perhaps due to the absence of an examiner and other speakers. There has been quite a lot of research into the impact on performance of interlocutor-related variables (for example, Porter 1991a, 1991b, Porter & Shen 1991, O'Sullivan 2000a, 2000b, 2002). One advantage of an interlocutor-free test such as the computer test described here is to remove this as a source of construct-irrelevant variance.

In general, candidates found that the interlocutor variables did not affect their performance in the test in a negative way, even if staff participants felt otherwise.

Table 5.14b

Document Analysis	
Document	*Findings*
CB test description/design	Interlocutor variables including examiner voice (task A & B) and other speakers (in task B) described in test specifications
CB test script	Evident here
Rating criteria (new TOEFL)	N/A

Commentary

The specifications for the test describe interlocutor variables in terms of each validity element, including the examiner in task A and B (test instructions and test tasks information) and the speakers in the dialogue in task B.

Table 5.14c

Observation	*Findings*
Researcher	Students did not seem to be affected by interlocutor variables, i.e. no questions were raised relating to speech rate, accent, and gender.

Commentary

In general, the students were able to fulfil the tasks without major problems relating to interlocutor variables.

Interlocutor variables for task B:

Interlocutor speech rate; accent; acquaintanceship; gender; interlocutor presence; turn-taking

Table 5.15a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	69% agreed **interlocutor speech rate** suitable 75% disagreed **accent** difficult 59% agreed **acquaintanceship** important 80% agreed **gender** not a problem 23% agreed prefer **same gender;** 25% disagreed; 52% no view 85% agreed that understanding **rules of turn-taking** is important
Questionnaire	Lecturer/Examiner (N= 29)	27% agreed **interlocutor speech rate** suitable; 12% disagreed 56% disagreed **accent** difficult 39% agreed **acquaintanceship** important 17% agreed **gender** not a problem; 37% uncertain 12% agreed prefer **same gender**; 42% uncertain 37% agreed **examiner not present** preferred; 20%uncertain 37% agreed that understanding **rules of turn-taking** is important; 7% disagreed; 15% uncertain
Questionnaire	Administrator (N = 7)	
Questionnaire	Expert (N = 9)	
Interview	Student (N = 13)	Mixed reactions; some still prefer to interact with real speakers as eye contact & gestures are important in communication
Interview (group)	Staff (N=30)	N/A

Commentary

Like data for task A above, staff responses showed lower percentages for the interlocutor variables in task B than student responses. In addition, there were missing data for staff responses for task B as well. Hence, 75% of students disagreed that accent is difficult for them, 56% of staff disagreed; 80% of students claimed gender is not a problem for them, only 17% agreed to this with 37% uncertainty. However, in task A students' disagreement for same gender examiner was higher (67%) than for task B (25%) with 52% who were uncertain. Perhaps because task B is different from task A in

terms of interlocutor presentation (with pictures and in dialogue), students were affected somewhat; again, this is the first time they had experienced such input for the test.

All this points to the fact that students may be affected by the examiner/interlocutor variables during the test. They need to be informed and made aware of the interlocutor(s) involved in the test. Some students also indicated that while the computer test is fairly easy and interesting, having real people to talk to is important for other aspects of communication such as eye contact and bodily gestures.

In general, students were not severely affected by the interlocutor variables for tasks B, even though lecturers/examiners might think otherwise.

Table 5.15b

Document Analysis	
Document	***Findings***
CB test description/design	Interlocutor variables including examiner voice (task A & B) and other speakers (in task B) described in test specifications
CB test script	Evident here
Rating criteria (new TOEFL)	N/A

Commentary

Like task A, the specifications for the test describe interlocutor variables in terms of each validity element, including the examiner for test instructions and test tasks information and the speakers in the dialogue in task B.

Table 5.15c

Observation	*Findings*
Researcher	Students did not seem to be affected by interlocutor variables, i.e. no questions were raised relating to speech rate, accent, and gender.

Commentary

Again, it is difficult to establish if students were in fact affected by the interlocutor variables in the computer test. However, observations indicated that they were not affected in a negative way, based on their conduct during the test; no questions or problems were raised in relation to these variables.

5.4.5 OVERALL COMMENT ON TASK DEMANDS

The main advantage of the computer test, standardization in terms of test input and administration, was detailed above. The task demand elements of the computer test are similar to those in the direct test and we mentioned above how channel of communication, topic/content knowledge, and nature of information are standardized for all candidates. The issue of equivalent forms of the test, especially since six sets are prepared for the direct test each term, is addressed by the computer test. In addition, elements of linguistic variables and examiner/interlocutor variable are also controlled in the computer test. It is interesting to note that our data for these elements have improved dramatically; for example, all participants agreed lexical and structural range of test instructions and test tasks were appropriate and suitable for the candidates' level (see table 5.13a above). These figures show a big increase from data for the same elements in the direct test (chapter 4, table 4.15a). In relation to interlocutor variables,

though staff members seemed sceptical, students were fairly confident of these elements; their responses relating to interlocutor speech rate, accent, acquaintanceship, and gender were positive. Comments made by students in relation to this, though mixed, were forthcoming. Some realized that doing the test using the computer is something new and will require time to get used to; others found it innovative, fairly easy and an effective way of encouraging those who are shy to speak more. Yet, others were intimidated in the sense that it was a new experience, and they are alone with the computer, even though their computer familiarity levels are high. (see Appendix 3I in CD 1 attached for data on computer familiarity)

Hence, it appears at this point that data indicates high context validity for the elements of the computer test. Mainly, the computer test is able to address the problems raised in the direct test study relating to test unfairness due to lack of equivalence in test forms, test reliability due to problems with co-construction of discourse in the group task, inconsistent test administration, and task demands which were unclear and often unpredictable for the candidates. The computer test is able to overcome these difficulties through standardization of test content, administration and input as well as output of task demands.

This next part of the findings is on theory-based validity, i.e. what and how test takers' did in order to attend to and fulfil the tasks in the computer test. We recap from chapters 3 and 4 that even though we describe and report these validity components separately, it is their interaction (context, theory-based and scoring criteria) that is at the crux of construct validity (Weir 2005). The intention at this point is not just to investigate

the candidate's strategies and thoughts for the computer test, but also to see how these might differ from those employed in the direct test. The elements covered for the computer test are the same as the ones in the direct test. Hence, the details of what theory-based validity is, its elements of internal processing (executive process & resource) are found in section 4.3.5 of chapter 4 on findings for theory-based validity of the direct test. (see Figure 5.4 below)

In this part, our source of data came from student participants (questionnaire and interview), and observations conducted by the researcher.

Figure 5.4 Aspects of Theory-based Validity for Speaking (from Weir, 2004)

THEORY-BASED VALIDITY		
INTERNAL PROCESSES • Conceptualiser • Pre verbal message • Linguistic formulator • Phonetic plan • Articulator • Overt speech • Audition • Speech comprehension	MONITORING	EXECUTIVE RESOURCES **Content knowledge** • Internal • External **Language knowledge** • Grammatical • Discoursal • Functional • Sociolinguistic

5.4.6 THEORY-BASED VALIDITY

5.4.6 a) Executive Process involves what candidates did for each task, for each of three stages: before starting the task, during planning time, while performing the task, and at each stage what processes are involved.

Task A

What I thought of/did before I started: ***Conceptualiser phase***

Table 5.16a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	94% agreed they **read carefully** 90% agreed they **thought of points** 92% agreed they **wrote down points** 61% agreed they thought of **how to satisfy examiners**
Interview	Student (N = 13)	• Some candidates think about examiners even though they were not present during the test

Commentary

The percentage of agreement on all strategies except 'thought of how to satisfy examiner' are high, a slight improvement from the data in the pilot study. They are however, similar to the data found for the direct test. Hence, most candidates go through similar processes before they even begin the test for both the direct and computer test; they employ some level of conceptualization at this stage. In addition, there was 61% agreement on the item relating to examiners; some students mentioned they thought of examiners even though the examiners were not there in person.

In general, students do go through some thought processes even before they begin the test and this seemed similar for both methods of testing.

Table 5.16b

Observation	*Findings*
Researcher	It was difficult to ascertain, but students seemed alert at the start of the test, listening to a brief introduction and instructions from the technical assistant

Commentary

Many hours of observations showed candidates to be attentive of the information provided before the test began in terms of a brief introduction and technical instructions for the test. It was difficult to establish if in fact some form of processing took place amongst them, but data above indicated some processing of the tasks assigned to them.

What I thought of/did before I started: ***Linguistic formulator/Phonetic plan/Grammatical***

Table 5.17a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	62% agreed they **thought of words and expressions** needed 71% agreed they **thought of structures** needed 57% agreed they **practiced speech in their minds**
Interview	Student (N = 13)	N/A

Commentary

As the table above shows, candidates went through some form of processing in thinking about the lexical and structural range they might need for the test, and more than half of them agreed that they practised the speech in their minds. These figures are again, similar to the figures found in our direct test data. Hence, students did go through similar processes for both tests at this early stage.

Table 5.17b

Observation	*Findings*
Researcher	Like the data above, it was difficult to ascertain, if students actually thought about lexical and structural range needed for the tasks, but they didn't raise questions relating to these elements, i.e. they seemed to have no problems here

Commentary

As with other aspects of cognitive processing, it is difficult to determine if candidates are indeed going through these processes, but our data and observations indicated they did think about words and expressions and structures needed for the tasks. In fact, test score data for both tests (see Appendix 3J in CD 1 attached for test scores from both tests) showed performances that were fairly well thought of and executed, i.e. candidates who indicated that they went through the process, scored quite well in the test (this will be discussed in detail in chapter 6 when data from both tests are compared and analysed).

What I thought of/did during planning time

Table 5.18a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	77% agreed they **thought of words and expressions** 59% agreed they **thought of structures** needed 72% agreed they **thought only in English** 33% agreed they **thought only in own language** 64% agreed they **thought both in English & own language** 85% **planned ideas for each topic in the mind** 56% disagreed/uncertain about **ability to put ideas in good order**
Interview	Student (N = 13)	Some students mentioned their inability to put ideas in good order

Commentary

It is curious to see that data for some elements above show a slight improvement from the pilot study data; for example, an increase for 'thought of structures needed' in both before starting and planning stages, and fewer candidates 'thought in one language' for this study compared to the pilot. It was also encouraging to note that the percentage for 'able to put ideas in good order' had increased in this study (previously below 20% able to put ideas in good order; now more than 40%). Another interesting point is in the high percentage of students who planned for ideas in their minds rather than writing notes, which happened a lot more in the direct test, possibly because there was more time for preparation.

In general, candidates appeared to go through several processes at this planning stage.

Table 5.18b

Observation	*Findings*
Researcher	Students appeared intent in their planning of the tasks, i.e. during the preparation time given for each task. Whether they thought about words, grammar, in which language, weren't clear; they did seem deep in thought and were occupied at the preparation stage

Commentary

As stated before, observing thought processes is a difficult task, and one cannot state with certainty what candidates are going through. However, data from table 5.16a, 5.17a and 5.18a above indicated that students did go through some degree of thought processes at the various stages of task A. We could say with some confidence that our

observations showed students utilizing the preparation times and were occupied during those times.

What I thought of/did while speaking: *Overt Speech (Presentation):*

Table 5.19a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	56% **checked** word use 48% checked grammatical accuracy 61% checked organization of presentation 79% **adjusted** word use 59% adjusted grammatical accuracy 74% adjusted organization of presentation
Interview	Student (N = 13)	Students indicated they have trouble with word use and organizing ideas

Commentary

The figures for 'checked appropriateness of ...' are lower than the figures for 'adjusted appropriateness of...' More interesting is the percentages for both 'checked organization of presentation' (previously 45%) and 'adjusted organization of presentation' (previously 47%) have increased from the pilot data and direct test data (see chapter 4, section 4.3.5, table 4.19a). This is possibly due to several reasons such as a clearer test purpose, sufficient external knowledge, and mainly because the absence of a live interlocutor makes self-repair and monitoring easier as the situation becomes less intimidating for the student; this is a key factor in the computer test.

Hence, it is apparent that candidates performed some level of self-monitoring at this stage where they intend to verbalize their ideas.

Table 5.19b

Observation	*Findings*
Researcher	What students do while they are speaking is hard to determine. Whether they monitor or not, students seemed to manage to speak through the computer without much trouble

Commentary

Even though observations were made of students speaking and recording their voices on the computer, it is difficult to determine what they actually do in terms of processing. However, students spoke within the time limits without much trouble in terms of completing the task, and our data above indicates that some level of monitoring takes place for these students when they adjusted their word use and grammar, and even organization of their presentations.

Task B

What I thought of/did before I started: ***Conceptualiser/Preverbal message***

Table 5.20a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	95% agreed they **read carefully** 93% agreed they **thought of points** 93% agreed they **wrote down points** 68% agreed they thought of **how to satisfy examiners**
Interview	Student (N = 13)	• Some candidates think about examiners even though they were not present during the test

Commentary

Like data for task A (table 5.16a above) students' agreement on the processes they went through even before the test started is in the 90s percentage; on satisfying the examiner, the percentage is also moderate. As shown in Levelt's speech process (1989), the initial stage of conceptualization is likely to be the most difficult for the speaker because at this stage the message is conceived and formulated.

Table 5.20b

Observation	*Findings*
Researcher	Again, what students do while they are preparing for the task is hard to determine as these processes are internal, and in the computer test, very individual and private.

Commentary

Internal processing is almost impossible to observe and so is the case with our students. However, their responses to the questionnaire items relating to the various stages of processing during the test provided us with an idea of what they went through.

In general, students carried out some level of mental preparation in terms of word use, grammar and organization of presentation, before the test began.

What I thought of/did before I started: ***Linguistic formulator/Phonetic plan/Grammatical***

Table 5.21a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	72% agreed they **thought of words and expressions** needed 75% agreed they **thought of structures** needed 72% agreed that **information from task A helped them prepare for task B**
Interview	Student (N = 13)	Students indicated they have trouble with word use

Commentary

In general, data for all items above show an increase from data in the pilot study, and similar to data in the direct test study. Our students have had practice for the direct test in their classes, and for this reason, they expressed more confidence for this element in the direct test, i.e. they thought of word use and grammar, and for task B, they know the format well and realized the importance of information in task A for task B. In spite of these factors, students went through the same processes for the computer test in terms of their linguistic formulator.

Hence, students' processing at this stage is in line with the speech process, and processes involved for the direct test.

Table 5.21b

Observation	*Findings*
Researcher	Students were engaged in some processing as they prepared for the test at this early stage

Commentary

During the test, students were obviously engaged in some level of preparation, be it mental or otherwise. Besides, for most of them, this is their first attempt at a computer-delivered speaking test, and this would certainly have an effect on processing. Either way, students were engaged in some level of processing at this stage.

What I thought of/did during planning time

Table 5.22a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	82% agreed they **thought of words and expressions** 66% agreed they **thought of structures** needed 72% agreed they **thought only in English** 31% agreed they **thought only in own language** 66% agreed they **thought both in English & own language** 84% **planned ideas for each topic in the mind** 54% agreed they were **able to put ideas in good order**
Interview	Student (N = 13)	Some students mentioned their inability to put ideas in good order

Commentary

For this set of data, fewer candidates 'thought only in their own language' compared to findings in the pilot study. One of the reasons for the inclusion of this item in the theory-based questionnaire is feedback received from staff members in the direct test study where they claimed students' performance was hindered by their constant recourse to their own language. The data obtained here shows that a low percentage of them do this at the planning stage. Hence, this may have been an overgeneralization on the part of lecturers.

In terms of 'able to put ideas in good order', the percentage is modest, is higher than the percentage in the pilot study, and similar to data for the direct test. This is not surprising when one considers organization of ideas as more difficult when the interlocutor is present as in the face-to-face test. It hinges again on the problem of co-construction of discourse since a speaker's response is highly dependent on what has been said before by another speaker.

In general, students were able to process various elements at the same time during the preparation stage of the test.

Table 5.22b

Observation	*Findings*
Researcher	Students were engaged in some processing as they prepared for the test at this early stage

Commentary

Similar to the preparation stage for task A, students were engaged in cognitive processing of the input from the test at the planning stage; they appeared to be able to process the information from the test whilst thinking through at least five of the elements above.

In general, students were capable of some level of cognitive processing during the preparation/planning stage of the test.

What I thought of/did while speaking: *Overt Speech (Presentation)*

Table 5.23a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	71% **checked** word use 53% checked grammatical accuracy 77% **adjusted** word use 56% adjusted grammatical accuracy 80% adjusted point(s) I want to make 74% **checked** the points they make when others are speaking 77% **adjusted** their response based on what was said
Interview	Student (N = 13)	Students indicated trouble with grammatical accuracy in their speech

Commentary

Data for this set of elements of validity are consistent with data from the pilot study, and shows improvement from the direct test data. The only low/modest percentage is related to 'grammatical accuracy' which has been consistent throughout the studies; students seemed to show low agreement when it came to thinking of, monitoring and making adjustments to grammar. An interesting finding is that even for the computer test, students indicated that they monitored and adjusted their ideas/points based on what was said. In this case interlocutor speech, i.e. two speakers engaged in a dialogue, who then engages the test taker into the dialogue with questions/prompts; the student has no choice but to be attentive of the dialogue that is in progress.

Hence, another advantage of the computer test is its ability to replicate task B of the direct test by engaging students are in the dialogue. Data from staff rating exercise of the computer test (see section 5.4.9 below on Scoring validity) highlighted this point;

raters noticed improvement in candidates' speech as they progressed from task A to several parts in task B.

Table 5.23b

Observation	*Findings*
Researcher	Students were engaged in their own processing as they speak during the test

Commentary

The data above is overwhelming of the evidence that students engage in some form and level of processing during their presentations and response to the dialogue in task B. Our observations showed that while they speak, they engage in internal processing, especially during task B; they had to listen carefully and respond to the prompts, and this goes on progressively throughout the task.

In general, students employ some level of processing at the final stage of verbalizing their ideas and thoughts; it is not a straightforward linear process as it incorporates many elements including lexical, syntactical, and morphological appropriacy and accuracy, as well as consideration for interlocutor speech.

5.4.7 CONCLUSION FOR EXECUTIVE PROCESS

In general, the findings above indicate that not only do students engage in internal processing in the same way that they did for the direct test (see tables 5.16a, 5.17a, 5.18a), they are also able to perform certain strategies better in the computer test (see tables 5.19a, 5.21a, 5.22a, 5.23a) especially for task B. Task A may seem straightforward in terms of the task demands, i.e. elaborating on ideas/information provided by the test task and processing limited informational functions, but data for overt speech showed

an improvement from the same elements in the direct test. One of the main reasons seemed to be that the absence of the examiner/interlocutor could make monitoring and articulating less stressful as the candidates felt less intimidated (see Appendix 3.10 for evidence from staff to support this). The same results were found for task B, which for these candidates would have been the first time being presented with such a format for an 'interactive' task; they not only adjusted lexical and grammatical use but also the points they wanted to make based on what they hear of the interlocutor in the test. Hence, one of the main features of the computer test, its ability to replicate the direct test while controlling for task demand and setting, is an advantage to our candidates.

The evidence above is overwhelming that candidates go through some form of cognitive processing, at some level, as described in Levelt (1989). This process is not linear and straightforward but in fact involves utilizing several elements at the same time, e.g. at the conceptualization stage while candidates are monitoring and filtering input for the task , the linguistic formulator is also activated for lexical, grammatical and morphological accuracy. In terms of monitoring, Levelt's model of the speech process shows the speaker monitoring their own words to make sure they do not make errors. Speakers also make spontaneous self-repairs as they attend to various aspects of the action they are performing (see Field 2003); they monitor the message/concept, the way it should be expressed to avoid ambiguity (discoursal), level of formality required by the context of discourse (sociolinguistic), lexical, syntax and morphological (grammatical). In fact, not all source of "the speaker's trouble" (Field, 2003: 190) are given equal attention, and in most cases the speakers experiences most difficulty at the conceptual level. In the data above, candidates indicated that they go through the

conceptualization phase during preparation time for the task while monitoring its content and linguistic demands, and before as well as at the point of articulation; they monitor and self-repair their own language use and content knowledge.

Therefore, it is clear at this point that candidates go though internal processing when attempting the test tasks, that these processes are not linear but multifaceted, and that most processes were similar to the ones they employed during the direct test. In addition, the computer test has the advantages of removing the interlocutor factor which makes self-repair easier and less stressful for the candidates, and it is able to replicate the direct test while controlling the effects that test setting and task demands might have on the candidates.

5.4.6 b) Executive Resource refers to knowledge (content and language) which students possess either from past experience or readings (internal), or knowledge which they obtain from the test itself.

Content knowledge:

Internal (background) knowledge

Table 5.24a

Participant Data		
Instrument	***Informant(s)***	***Observation(s)***
Questionnaire	Student (N= 61)	64% agreed **topic** was familiar from previous readings and experience 56% agreed **information for the task** was familiar from previous readings and experience; 10% disagreed; 34% uncertain
Interview	Student (N = 13)	Topic fairly easy

Commentary

These data, though modest, indicate that the topic and information found in the tasks were appropriate and approximate the level of these candidates. In the direct test study, the percentages for these elements were slightly lower. It is not surprising that candidates were uncertain about information found in the computer tasks as this is the first time such information appeared in this manner.

In general, students realized the importance of internal knowledge as part of processing, without which they may fail the task in terms of content knowledge.

Table 5.24b

Observation	*Findings*
Researcher	Difficult to ascertain this aspect as well, even with data such as interview and test scores

Commentary

A student's level of background knowledge is difficult to determine through mere observation. We could however, control this aspect for all students by providing them with topics and information familiar to them and within their experience. For the computer test, the topic and content were selected from past papers of the direct test, hence appropriate for the candidates. It was observed that students were not negatively affected by the topic and test content, i.e. no questions were raised relating to these elements.

External knowledge

Table 5.25a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	**Task A**: 87% agreed that information in the **instructions** were necessary to complete task A **Task B**: 94% agreed that information in the **instructions** were necessary to complete task B
Interview	Student (N = 13)	N/A

Commentary

External knowledge refers to information provided by the test for the candidates. Students in this data set showed high agreement on this aspect of the test. As mentioned in 5.24a above, this aspect of the test can be controlled so that students are not disadvantaged and are able to fulfil the tasks adequately Hence, this aspect of the test has high validity if the instructions are equal and fair to all, and students are able to accomplish the task without seeking further clarification.

Table 5.25b

Observation	*Findings*
Researcher	Students appeared to be able to follow the instructions without request for further clarification or explanation

Commentary

It was observed that students were able to complete the tasks without further help relating to test instructions. We could confidently say that the instructions were appropriate and sufficient for students to fulfil the task.

Language knowledge for the computer test includes functional and sociolinguistic knowledge

Functional knowledge

Table 5.26a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	59% agreed they were able to connect what they said to what's been said 67% agreed they were able to conclude the discussion
Interview	Student (N = 13)	Concluding the task was a problem, especially due to insufficient time

Commentary

The data set for these candidates differ from those in the pilot study, especially for the first element; while this is a different set of candidates, all other aspects of the test is the same. In general, these students' response had shown an increase in most validity aspects, until this one. It is also noted that out of those who disagreed to being able to 'connect what they said to what has been said', 28% were uncertain. Once more, the absence of live interlocutors may have affected these students, even though table 5.23a above indicated otherwise, i.e. they were able to adjust their ideas/points based on what other speakers (on the computer) said.

The non-linear nature of the speech process is brought to light again; that overt speech is the outcome of a series of processes in varying combination and degree of attention, including the speaker's ability to demonstrate his/her functional knowledge to be able to connect to what others said and to conclude the discussion.

Table 5.26b

Observation	*Findings*
Researcher	Students seemed to respond to the prompts directed at them appropriately

Commentary

Although our data showed modest percentages, students seemed to be able to respond to the prompts directed at them appropriately, without the need for help or extra time.

Sociolinguistics knowledge

Table 5.27a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Student (N= 61)	79% agreement for task A 74% agreement for task B
Interview	Student (N = 13)	More of a question of formality, not so much language difference

Commentary

Perhaps it is confusing that this item was given in the questionnaire in relation to both tasks; knowing this difference shows some level of sociolinguistic knowledge, regardless of the context. It is not surprising that there were comments relating to formality rather than appropriate linguistic forms; in fact sociolinguistic knowledge encompass knowledge of appropriate linguistic forms and conventions characteristic of particular sociolinguistic groups, and the implications of their use, or non-use, such as slang, idiomatic expressions, dialects, cultural references, figures of speech, levels of

formality and registers'. In general, students indicated an understanding of this knowledge and realized its importance in relation to the test tasks.

Table 5.27b

Observation	*Findings*
Researcher	Students performed both tasks accordingly, without problems relating to the sociolinguistic knowledge required to fulfil the tasks

Commentary

Again, to determine if students have this knowledge, we would need to listen to their presentations and/or look at their test scores to detect if they had responded to the task according to its sociolinguistic demands. Our observations however, showed that students performed both tasks accordingly, without problems relating to the sociolinguistic knowledge required to fulfil the tasks.

5.4.8 CONCLUSION FOR EXECUTIVE RESOURCE

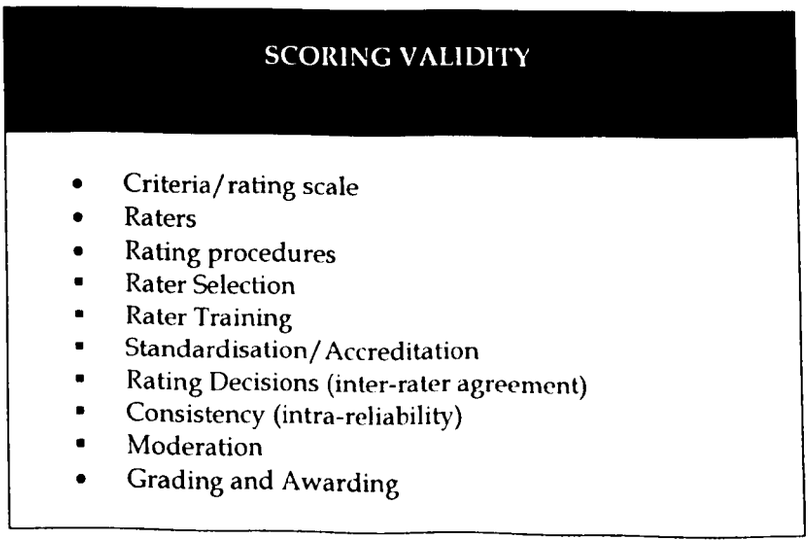
In terms of executive resources for the test, students regarded their content and language knowledge to be moderate, though they recognized the significance of these characteristics to succeed in the test (see tables 5.26a, 5.27a, above). That our candidates are aware of their own abilities (processing) and knowledge (content and language) is evident in the findings. As detailed above at the beginning of this chapter, the speech process involves not only the interaction between the test and the test taker, but also within the test taker's processing system. Hence, the executive resource is as important as the process itself for the success or failure of the task; processing is not effective if candidates are able to fulfil the task with total reliance on external knowledge and

without recourse to internal knowledge. If this were the case, not only context validity but also theory-based validity of the test is jeopardized.

While we present the framework for scoring validity below (Figure 5.5), we note that scoring validity for the computer test is almost non-existent at this stage. However, at the workshop conducted in May 2005, (Appendix 3.10) data was gathered from members of staff on rating of the computer test; rating the test using the old (UiTM) scale and the new (TOEFL) scale, and other issues on scoring validity. These findings are reported below.

For scoring validity of the computer test, staff participants were given the questionnaire with only three items relating to the criteria/rating scale. Other elements in the framework do not apply here since the test had never been used before, and this is the first time members of staff had seen the test. The elements of scoring validity are found in Figure 5.5 below.

Figure 5.5 Aspects of Scoring Validity for Speaking (from Weir, 2004)



5.4.9 SCORING VALIDITY for the computer test was limited to a set of rating criteria/scale for rating the test and its measures of consistency

Criteria/Rating scale Participants were asked if the components (Language use, Delivery, Topic development) covered all aspects of performance for the test, if they were sufficient to make a fair judgment and if they were clear to all markers.

Table 5.28a

Participant Data		
Instrument	*Informant(s)*	*Observation(s)*
Questionnaire	Lecturer/Examiner (N=25) Administrator (N = 7) Expert (N = 9)	1. Whether **components** (LU,D, TD) cover all aspects of the performance: 78% agreement for task A 56% agreement for task B 2. Whether three criteria **sufficient** for fair judgment: 54% agreement 3. Whether criteria are **clear** to all markers: 66% agreement
Interview		• TOEFL scale is detailed & with clear descriptions • Reliability of scoring (e.g. pass/fail cases can be easily identified) • There is a permanent record; tests can be re-played & re-evaluated • Potential for standardization across the board, especially across campuses

Commentary

The data above relate to the new TOEFL rating scale, which was introduced to staff at the workshop in May 2005 where lecturers, examiners and administrators used the scale to rate the computer test. All comments were gathered relating to the new scale. It is not surprising that the figures are modest, especially in terms of familiarity, even though it is higher than the results for the old scale which staff were familiar with and had used for many years (see chapter 4, section 4.3.9, table 4.25a). It was clear from staff members that the old scale was vague & not sufficient for rating the test. The new TOEFL scale

seemed more promising and even though participants had never seen or used it before and had no prior training in using it, data from the workshop suggested that the new scale was clearer and more effective for the markers when they used both scales to rate the direct test (data on scoring validity from Appendix 3.10). The comments above reflect the advantages of the computer test where scoring validity is concerned.

In general, it appeared that participants found the new scale to be constructive and that the computer test has great potentials in terms of reliability of scoring and standardization of scoring across campuses; these were major problems raised in scoring validity of the direct test.

Table 5.28b

Document Analysis	
Document	***Findings***
CB test description/design	Specifications for test include the scoring aspect
CB test script	Criteria listed and explained in the test
Rating criteria (new TOEFL)	Document from ETS site detailing criteria

Commentary

Documentation for scoring validity elements is found in the test specifications, criteria are listed in the test script for students and markers, and the new rating scale is available on ETS website, http://www.ets.org (locate in Download library: Scoring Guide (Rubrics) for Speaking responses).

Table 5.28c

Observation	*Findings*
Researcher	During the workshop, participants discussed and used the new TOEFL scale in spite of their first exposure to it and, no previous training on how to use it.

Commentary

Observations of the discussion on scoring validity showed participants' interest in the new TOEFL scale. The rating exercise that staff members participated in, in which they used the old and new scale to rate the direct test, then the computer test, produced encouraging results (Appendix 3.10). Lecturers, examiners and administrators were concerned about the development and content of the old scale, as compared to the new scale.

Table 5.29 ***Rating Decisions (inter-rater agreement)***

As stated for table 5.28a above, the outcome from rating the direct test using both the old (UiTM) and the new (TOEFL) scales at the May 2005 workshop showed promising results from the use of the new scale. For the candidates in Main Study 2 who took the computer test, the rater agreement was 0.8; this is a positive correlation in spite of the fact that raters had never used it before and had no prior training in using it.

Table 5.30 ***Consistency (intra-reliability)***

The overall correlation between rating of the computer test and the direct test for Main Study 2 (same candidates had taken both tests, same raters) was 0.7; hence a positive correlation again.

(**Note:** For detailed outcome of the above results, see Appendix 3K in CD I attached)

5.4.10 CONCLUSION FOR SCORING VALIDITY

Even though data for this section of the test are limited because of reasons cited above in the introduction, we can state at this point that response from participants regarding the scoring validity of the computer test is reasonable. This is mainly due to data from the direct test study (see chapter 4) which showed an overwhelming concern regarding the old (UiTM) scale and other factors related to scoring validity such as rater training, standardization and so on. The old rating scale was inefficient and difficult to use, and documentation of its development was not available for members of staff to refer to. Understandably, staff showed concern about the new scale but in effect, the new TOEFL scale, though rather detailed, wordy & can be difficult to use (as it was developed for commercial use), is a good starting point for developing a more balanced scale for UiTM. It is well-researched, is used for an established exam, has detailed descriptions of characteristics for each band, and is holistic in nature (practicality).

More importantly, participants were able to realize the potentials of the computer test where scoring validity is concerned such as:

- Reliability of scoring increases, for e.g. pass/fail cases can be easily identified using the scale, rating is conducted later so inter-rater reliability problem is removed, and there is a permanent record so tests can be re-played & re-evaluated
- Potential for standardization across the board, especially across campuses

Additional data from this study indicated high rater agreement between raters for the computer test, and for the overall correlation between rating of the two tests. While these results may point to the rating for this group of participants in this study, they are nevertheless positive, and the outcome from rating sessions using a new scale had only produced encouraging results.

5.5 SUMMARY OF THE VALIDITY OF THE COMPUTER TEST

At the beginning of this chapter, we stated the research question that this chapter relates to, i.e. to ascertain if a proposed semi direct web-based speaking test is valid in terms of context validity, theory-based validity and scoring validity. We also summarized the limitations of the direct test, which were discovered from the validation study conducted on that test. Hence, the main purpose of determining the validity of the computer test is clear: to address the research question, and so that the data can be used to address the target of the study, which is to overcome the problems faced by the direct speaking test using the computer as an alternative method of delivery.

One of the major findings of the direct test study was its design. Although the contextual aspects of the computer test were parallel to the direct test, our data above showed the main advantage of the computer test, i.e. one of standardization. The test is able to standardize elements of test setting (especially purpose, response format, order of items, and time constraint) and task demands (especially length, nature of information, topic, lexical range, structural range, functional range and interlocutor variables). This is crucial as we deal with the problem of equivalence of input in the

direct test; too much variance in test context results in construct irrelevance and students are disadvantaged.

Another related aspect is in standardization of test administration which is the main feature of the computer test. Consequently, students will receive a test that has been standardized in all these aspects and test validity increases. In terms of theory-based validity, students' responses indicated that the processing involved in attempting the computer test is similar to, in some cases (see tables 5.19a, 5.21a, 5.22a, 5.23a above) better than, the processing of the direct test. This means that the computer test not only approximates the direct test, it is an improvement in terms of context validity and internal processing is not severely affected. Our students seemed to be able to go through the tasks accordingly; they went through the various stages of conceptualizing the input, formulating their presentations, monitoring, and performing self-repairs as they spoke through the test.

Finally, while scoring validity needs to be investigated further, initial data and feedback from staff members indicated agreement on the potential use of a new rating scale, and that the computer test has advantages in terms of scoring validity.

The table below (Figure 5.6) summarizes the findings relating to context validity, theory-based validity and scoring validity of the computer delivered speaking test.

Figure 5.6 Summary of validity of the computer test

Validity Component	Validity Element	Positive finding	Negative finding	Mixed finding
CONTEXT VALIDITY	Purpose	♦		
	Response format	♦		♦
	Weighting			
	Known criteria	♦		
	Order of items	♦		
	Time constraint	♦		
	Physical conditions	♦		
	Uniformity of administration	♦		
	Security		♦	
	Nature of information	♦		
	Content knowledge required	♦		
	Lexical range	♦		
	Structural range	♦		
	Functional range	♦		
	Interlocutor variables	♦		
THEORY-BASED VALIDITY	Conceptualiser	♦		
	Linguistic formulator	♦		
	Overt Speech	♦		
	Internal knowledge		♦	
	External knowledge	♦		
	Functional knowledge	♦		
	Sociolinguistics knowledge	♦		
SCORING VALIDITY	Criteria/Rating scale	♦		
	Rating Decisions (inter-rater agreement)	♦		
	Consistency (intra-reliability)	♦		

Hence, the question of whether the computer test addresses the problems faced by the direct test is in partly answered. Even though a lot more data can be gathered on all aspects of validity, these initial findings show positive signs and potentials of the computer test.

Chapter 6: DISCUSSION

A comparison of findings from the direct test & the computer test

6.1 INTRODUCTION

In this chapter we compare the findings from the direct test (chapter 4) and the computer test (chapter 5) validation studies in order to address the final research question of the study, which is:

Can we justify replacing a direct test with a semi direct test of speaking ability?

Before we discuss further, it is important to re-visit the objectives of the study at this point. As detailed in chapter 1, the objective of the study was to investigate the validity of testing spoken language using computer technology as an attempt to overcome the issues of validity, reliability and efficiency of a large-scale speaking test administered to thousands of students across campuses of Mara University of Technology in Malaysia. To achieve this, a validation study was conducted on the direct test, its outcome and findings incorporated into the development of a computer test, which was then subjected to various investigations of its own in terms of validity and reliability. However, in past chapters, several significant points had been constantly alluded to because they are essential to the design and methodology of the present study.

Firstly, the framework used for validating a speaking test is new (Weir 2004/2005); it was employed as it is by far the most comprehensive framework of its kind in the literature, (see chapter 2, section 3.1 for details) and this enabled the researcher to gather

data in a systematic and objective manner. In fact, this is the first attempt at operationalizing the framework for the purpose of test validation. The current study however, focused on three components, i.e. context, theory-based and scoring validity. According to Weir (2005), the interaction between context validity and theory-based validity, and the interaction between them and the scoring criteria, is at the heart of construct validity. This is the main intention of the validation studies for the direct test and computer test, which is to ensure the test is measuring what it sets out to measure in terms of validity and reliability, and to enable a comparison between them. Only then are we able to address the research question above.

Secondly, it was the first time such a validation exercise had been conducted on the direct test at the university, and therefore, it was also the first time all the participants involved in the study had taken part in such a study. Nonetheless, as reported in chapter 4, data gathered on the direct test were overwhelming and revealing of the status of the test. Furthermore, students were then exposed to the speaking test through the computer, also a first time experience for them. Once more, the data, though fewer in terms of number, were encouraging in terms of the prospects of using the computer for a speaking test of this nature.

The fact that the computer test paralleled the existing test in most aspects was also mentioned repeatedly in past chapters. Not only was this a key factor in the design of the computer test, but data from the direct test validation study provided further contributory elements to the development of the test; the problems arising from the direct test had to be addressed by the computer test. Hence, the computer test was

developed from specific objectives of the main study, and more importantly, its design was driven by the outcome of the direct test study. This element of parallelism between the tests also facilitates their comparison in the present chapter.

At this juncture, it is evident that many aspects of the current study, especially its design and methodological considerations, were innovative and conducted for the first time. We can now compare the outcomes of these two investigations to determine if there are indeed similarities or differences between the tests, to answer the research question.

Figure 6.1 below shows the outcomes of both direct test and computer test studies in terms of their findings (positive, negative, mixed). We recap that for both studies data were gathered through questionnaire and interview (formal & informal) techniques, observation of test sessions, and an analysis of test documents.

This chapter makes the comparison of these findings according to the test taker/theory-based validity, the test/context validity, and scoring validity. Additional data was gathered from staff members at a workshop in May 2005 relating to rating criteria/scale, and from the direct test and computer test scores.

Figure 6.1 A comparison of findings for the direct test (from chapter 4: Main Study 1) and the computer test (from chapter 5: Main Study 2)

		Direct Test			Computer Test		
Validity Component	**Validity Element**	**Positive finding**	**Negative finding**	**Mixed finding**	**Positive finding**	**Negative finding**	**Mixed finding**
CONTEXT VALIDITY	Purpose	♦			♦		
	Response format			♦	♦		
	Weighting		♦				♦
	Known criteria		♦		♦		
	Order of items			♦	♦		
	Time constraint		♦				♦
	Physical conditions		♦		♦		
	Uniformity of administration		♦		♦		
	Security		♦			♦	
	Channel		♦		♦		
	Discourse mode			♦			
	Test length		♦		♦		
	Nature of information		♦		♦		
	Content knowledge		♦		♦		
	Lexical range	♦			♦		
	Structural range		♦		♦		
	Functional range		♦		♦		
	Interlocutor variables		♦		♦		
THEORY-BASED VALIDITY	Conceptualiser			♦	♦		
	Linguistic formulator			♦	♦		
	Overt Speech			♦	♦		
	Internal knowledge		♦				♦
	External knowledge		♦		♦		
	Grammatical knowledge		♦				
	Discoursal knowledge		♦				
	Functional knowledge			♦	♦		
	Sociolinguistics knowledge			♦	♦		
SCORING VALIDITY	Criteria/Rating scale		♦		♦		
	Rater training		♦				
	Standardization		♦				
	Rating condition		♦				
	Moderation			♦			
	Statistical analysis		♦				
	Grading & awarding			♦			
	Rating Decisions (inter-rater agreement)	♦			♦		
	Consistency (intra-reliability)	♦			♦		

It is evident from the table that the direct test has more bullets in the negative findings column than the computer test. This is found in all components; although the computer test has only one element under scoring validity, the outcome is mainly positive.

6.2 COMPARISON OF FINDINGS ACCORDING TO THE RESPECTIVE VALIDITY COMPONENTS.

A. Context validity

In chapter 4 it was concluded that the direct test lacked context validity for various reasons. One of them is related to members of staff involved in developing the test (lecturers, examiners, and administrators) who lacked clear knowledge on issues related to testing spoken language and factors that affect or drive performance. When lecturers lack the necessary knowledge, students become disadvantaged as they were not well informed regarding test demands as well as internal processing required of them to fulfill the test. In addition, our findings showed inconsistencies in terms of the information that students received for the test, such as for the criteria used for rating their performance, nature of information in the test content, and interlocutor variables. In fact, part of the problem is related to test documentation, which lacked sufficient and appropriate content. The test specifications and other related documents did not equip lecturers/examiners with adequate information. Overall, the test did not fulfill the criteria for context validity of the validation framework, i.e. it was not clear if the test was created based on a theoretical basis/model. When documents are incomplete and test specifications inadequate, lecturers, examiners, and students alike are not able to obtain pertinent information relating to the test effectively and accessibly. As a result, lecturers are ill equipped with relevant information that they need to pass on to students; they carry on with information they know from experience, rather than from clear specifications of the test.

More essentially, because the test was not driven by a model of language learning/ testing, it fell short of its objectives. It shows lack of knowledge regarding students' cognitive and meta-cognitive processing for such a test, test design becomes invalid, and reliability of test scores becomes questionable. The crucial point here is that the features of test tasks have the potential to affect test reliability (Hughes 2003); this is directly related to scoring criteria, which were decided upon based on the tasks, and the concern is whether they match the construct (see section 'Scoring validity' below for further discussion on test reliability). In terms of validity, it had been reiterated that the components are interconnected; context validity impacts on scoring validity as well as theory-based validity. The findings in chapter 4 is evident of this fact: a clear purpose affects goal setting and monitoring, response format affects processing, hence rating, planning time affects goal setting, and even knowing the criteria for rating influences how a candidate approaches the task at hand.

In chapter 5, it was concluded that overall, context validity for the computer test looked more assuring. While the direct test lacked a clear theoretical basis for its design and development, literature on the computer test was reviewed and its potential in overcoming the problems arising from the direct test was explored (see chapter 2, section 2.123 for details on computerized testing; section 3.1 on validation and how this addresses problems in the speaking test). In addition, feedback gathered from students who were directly involved in the computer trials and staff who had seen it and were involved in the rating of the test, were overwhelming and positive in most aspects of test context. Furthermore, in terms of scoring criteria, a new rating scale was introduced

and proposed; it had been well researched, has detailed descriptors, is used by a major exam board, and has the practical advantage of a holistic scale. (see discussion below on 'Scoring validity').

With computerization, we can include features, which could address context validity. For example, it is now possible to select the language for instructions/test rubrics, as it is crucial that test instructions should not be more difficult than the text or task (see DIALANG, http://www.dialang.org for examples). We are also able to control test purpose and information such that candidates receive equal input; no candidate is disadvantaged from inconsistencies in topic, nature of information, test length and so on. In terms of response format, the format we choose can critically affect a candidate's cognitive processing of the task, i.e. it will affect theory-based validity. Alderson (1995) asserts that since the effects of response formats tend to be unpredictable, it can be a potential source of construct-irrelevant variance; hence more than one response format should be used to test any ability. While this is relevant for the speaking test, it is not common practise in classroom based testing due to practical concerns. More importantly, the more formats, interlocutors, topics, and time we provide for the test taker, the more variances we get in terms of the characteristics and administration of these elements in the test; this will inevitably affect performance. With the computer test, this can be overcome; more formats can be used either at one time or over several sessions without jeopardizing test purpose and objectives.

We can deduce from the above that the computer-delivered speaking test has a clear advantage over the direct speaking test. One of its major benefits is its ability to

standardize the elements which appeared to cause serious problems for the test taker in the direct test. These include a clear purpose, response format, time constraint, topic/content knowledge, nature of information, linguistic variables, interlocutor variables and test administration. The standardization of these elements addresses the concerns of equivalence of multiple forms of the test, variability in test administration, and co-construction of discourse in interaction (task B). Luoma (2004: 37) points out that in an interactive task, the examinee's talk is almost inevitably influenced by other participants' personality, communicative style and possibly language level. The main concern here is that all test takers may not get an equal chance to show their best speaking skills and ability (see Weir 1993; Iwashita 1999). Fulcher (2003: 44-46) provides a detailed account of what it means when performances are co-constructed, based on the works of McNamara (1997), Vygotsky theory of language ability, the Interactional Competence theory (ICT) (Kramsch 1986), He & Young (1998) and Young (2002), among others. He raised the important questions of how scores can be given to an individual when it is in fact a paired or group task, and how the contributions of the interlocutor/rater can be taken into account when he/she is co-responsible for the construction of talk. These crucial matters do not arise in the computer test; in our test, the nature of task B is such that it encourages a certain degree of interactive performance in the form of responding to comments/questions driven by the context, but controlled for all candidates. Though more work needs to be undertaken for the inclusion of such a task in the test, the results so far are positive. The fact that earlier examinations of the speech functions elicited by candidates for the group discussion showed a mismatch from test objectives, i.e. there was minimal or no demonstration of

interactional functions and no agenda management is evidence of the complexity of testing, and rating, such activities.

Thus, the main advantage of the computer test over the direct test is in standardization. This ability enables the test to address the concern for test fairness arising from test setting, task demands (input & output) and test administration, which are not equivalent and often not made explicit to the test taker. We are able to control the essential features of the test and this helps reduce variability in terms of performance, which in turn improves rating.

More importantly, while this is a first attempt at such a method of testing general spoken language ability, it offers an alternative to the traditional method without jeopardizing validity and reliability of the test. We can conclude that in terms of context validity, the computer test shows more positive results than the direct test in terms of all elements except security and internal knowledge; two items which were also negative findings in the direct test.

B. Theory-based validity

For theory-based validity, there seemed to be a conflict in the data from student participants and members of staff. Taking into consideration that this aspect of validity relates to assumptions that test developers, writers, experts etc. make about students' ability and knowledge of spoken language, what processes they should employ and knowledge they should have (internal, external, linguistic) to be able to perform the tasks well, we find that data from staff are contradictory to the students' claims of their

interactional functions and no agenda management is evidence of the complexity of testing, and rating, such activities.

Thus, the main advantage of the computer test over the direct test is in standardization. This ability enables the test to address the concern for test fairness arising from test setting, task demands (input & output) and test administration, which are not equivalent and often not made explicit to the test taker. We are able to control the essential features of the test and this helps reduce variability in terms of performance, which in turn improves rating.

More importantly, while this is a first attempt at such a method of testing general spoken language ability, it offers an alternative to the traditional method without jeopardizing validity and reliability of the test. We can conclude that in terms of context validity, the computer test shows more positive results than the direct test in terms of all elements except security and internal knowledge; two items which were also negative findings in the direct test.

B. Theory-based validity

For theory-based validity, there seemed to be a conflict in the data from student participants and members of staff. Taking into consideration that this aspect of validity relates to assumptions that test developers, writers, experts etc. make about students' ability and knowledge of spoken language, what processes they should employ and knowledge they should have (internal, external, linguistic) to be able to perform the tasks well, we find that data from staff are contradictory to the students' claims of their

ability to perform the tasks. Furthermore, that context validity affects the test taker in terms of his/her cognitive and meta-cognitive processing of the tasks, is demonstrated in this section of the findings of the direct test.

Evidence from the findings showed students' awareness of their strategies and internal processing of the tasks, albeit expressed through the questionnaires and interviews. For example, they were conscious of their actions at the preparation stage, of lexical and structural concerns, and organization of ideas. During the presentation, they were conscious of the same elements though constrained by time and presence of interlocutors; during interaction, they monitored their own speech as well as the speech of other speakers and indicated awareness of how their responses were influenced, in part, by what others say (discourse is co-constructed). On the contrary, lecturer/ examiners were of the opinion that in general, students were not able to perform the tasks successfully, let alone be aware of these processes and strategies. It is clear that while the test context affects student performance, an analyses of data gathered from other sources, including staff members, indicated inconsistent and inadequate information relating to how elements of the test influence processing, and how this then affects rating. Essentially, when the test itself lacks context validity due to reasons listed in the previous section, student performances are seen to be deficient and generally not satisfactory by test developers themselves. The situation here is perplexing; people involved in developing and administering the test did not demonstrate a clear perspective of the connection between their test and the test takers, and in turn decided the performances were below standards.

It is crucial that test developers make clear decisions about what operations they wish to be called into play by test tasks, and that these are made clear to the test takers. Weir (2005) asserts that construct definition is related to the particular types of activity that the examinees are asked to perform. He referred to Bygate's (1987) description of how speakers organize in 'routines' what they have to communicate as a good starting point. The repertoire of knowledge that learners have regarding informational and interactional routines reflect their familiarity with certain kinds of communication. Learners also need to be aware of 'improvisational skills' (Bygate's term) which are necessary for when interaction falters (see also Hughes, 2002: 'strategic competence' for when there is potential for communication breakdown). Other componential views of how interaction takes place and the processes of candidates include Levelt (1993), O'Loughlin (2001), and Hughes (2002). It is clear from the literature, though limited in terms of attention paid to the speaking process, that the more knowledge test takers have regarding elements of the speech process and interaction, the better their performance in the test will be. Students who are sufficiently grounded through previous experience regarding the speech routines, will have this knowledge readily available to them at the conceptualization stage in a spoken interaction; they will be able to select from a known repertoire at the planning stage. In terms of theory-based validity, these students have benefited from sound knowledge and processing becomes less burdensome. Weir (2005b) affirmed that evidence of the prime importance of preparing students well for the test, though limited, is out there. Students must be taught skills such as goal-setting, planning and monitoring, and time for these to happen must be provided in the test. In the case of the speaking test in the current study, the

students either had little or no knowledge at all regarding cognitive processing of test tasks.

With reference to executive resources, candidates for the direct test also failed to understand the linguistic knowledge required to fulfil the tasks efficiently, mainly due to lack of explicit information available to them from lecturers and in the test documents. It should be clear that for processing to take place adequately, language knowledge that candidates bring to the test must fit in and interact with those linguistic variables demanded of them by the task; this aspect of the test text/information should match those of the test taker's. Once more, if test developers' are inaccurate about the assumptions they make regarding the knowledge they should have (internal, external, linguistic) to be able to perform the tasks well, students' performance can be negatively affected. Our findings showed that in general, students were given insufficient information (or for some, no information at all) relating to the linguistic parameters of the test, both in terms of input and output. Limited work is found in the effect of these variables on the input/output for the speaking test; however, Alderson (2000) reminds us that while the task of identifying text variables that consistently cause difficulty may be complex, the interaction among syntactic, lexical, discourse and topic variables is such that no one variable is shown to be dominant (from Weir 2005, p 79). The job of test developers is to ensure that lexical, structural and functional variables in the test, both in input text and required as output, is appropriate for the level of the candidates, and that they are given equal emphasis in the test.

With reference to the computer test, Weir (2005: 144) cautions that the more indirect the task, the more difficult it will be to translate test results into statements about what candidates can or cannot do in terms of the target performance under review. He further illustrates the benefits of tasks such as peer interaction as representing real communication in terms of enabling candidates to cover both informational and interactional routines, reciprocity, and its replicability. However, without proper care especially at the test development phase, these properties could also cause potential problems for the test; context and theory-based validity of the test suffer. Candidates may well not be able to demonstrate the range of routines, including improvisational skills, if they have no knowledge or experience of these skills. Reciprocity is affected if one of the participants dominate the interaction, or if there is a large difference in proficiency between the candidates. Further problems may arise if one of the candidates is more interested in the topic or task, resulting in a one-sided interaction. In relation to co-construction of discourse, the issue of variability in input across candidates is one in which examiners have little control over, and inevitably affects rating and rater reliability.

How all this is related to the computer test is a matter which requires intensive and extensive attention, and the current study is one such attempt. Our table above indicates that in terms of theory-based validity, candidates seemed positive about their strategic and linguistic competence of the test. In chapter 5, we concluded that not only were students able to engage in internal processing in the same way that they did for the direct test, they were also able to perform certain strategies better in the computer test. This includes thinking through and adjusting word and expressions to use, grammatical

accuracy, organization of ideas, and responding to the questions/prompts appropriately, especially for task B. This appears to have been mainly due to the absence of the examiner and/or live interlocutor. Evidence given by the lecturers suggested that removing the interlocutor factor seemed to make self-repair easier and less stressful for the candidates.

Candidates indicated that they were more able to engage in the computer test tasks, reporting that they were more likely to engage in aspects of speech processing as described in Levelt (1989). It appeared that they were able to adjust lexical and grammatical use, and the points they wanted to make based on what they hear of the interlocutor in the test. Many hours of conducting the test and observations of test sessions, showed students having minimal trouble with test context and actual performance, indicated by lack of need for assistance; much of the assistance was needed for technical reasons such as recording and saving files.

Hence, the main advantage of the computer is its ability to replicate the direct test while controlling the effects that test setting and task demands might have on candidates' executive processes and processing. More importantly, the computer is able to address the issues mentioned above regarding interaction in speaking. As explained in previous chapters (chapter 2, 3, 4), it was ascertained, by using the observation checklist of speech functions (see O'Sullivan et al, 2002), that students in the UiTM test demonstrated informational routines and functions, with minimal or no interactional functions, and no agenda management. While it could be the case with this particular speaking test, literature on testing of speaking echoes the same concern for most interactive tasks. (see

Cohen & Olshtain 1993; Riggenbach 1998; Kormos 1999; Hughes 1998, 2002; Tannen 1989, 1996; Gass & Mackey 2000 and Mackey, Gass & McDonough 2000; Porter 1991a; McNamara 1996,1997 and O'Sullivan 2002). In addition, the students in the study were ESL speakers, at the intermediate/advanced level of the course, of the same age group, and had gone through the same number of hours for the speaking component. Nonetheless, unless students have been explicitly taught and have sufficient practice on these skills and routines, they will not be able demonstrate their best ability in the test.

The emerging concern for co-construction of discourse in interaction in the literature indicates its critical position in research related to testing spoken interaction. (see Brown (2003); Swain (2001); McNamara (1997); Fulcher (2003); Luoma (2004); and Dimitrova-Galaczi (2004)). Given the above matters, which are serious concerns for the direct test, the computer test appears to be a step in the right direction. In the test, more functions can be built in for candidates to demonstrate and recognise, and co-construction of discourse is eliminated as a concern. For example, in task B of the computer test, interactional functions of agreeing/disagreeing, justifying an opinion, making a decision, and summarizing, were built into the dialogue/test script; students may not experience interaction, but they should recognize the functions when evident in the task.

Hence, thus far, data from the computer test study has been encouraging. Students seemed to understand the tasks well, and were able to perform with minimal assistance and stress. We conclude from the study that context validity of the speaking test was

enhanced when delivered by the computer, and the impact on theory-based validity is positive.

C. Scoring validity

As stated above in context validity, it is essential that elements of test context are considered closely at the design and development stage as they could potentially affect the reliability of the test. Hughes (2003) made the point in relation to features of the test tasks in terms of the way they have the potential to affect test reliability. As a test is constructed, a mark scheme is drawn up, and as each task is developed in a speaking test, decisions are made on how it is marked, on weighting, and so on. Hence, the appropriateness of the marking scheme in relation to the task is crucial.

An illustration of this point was made in chapter 4, and is presented there in Figure 4.4. It shows how the underlying model or theory of ability, which steers a test, is directly linked to test tasks and the rating scale, which themselves affect rating, performance and the eventual meaning that we make of the test scores.

Without a proper understanding of this relationship, test developers and other parties connected to the test would not be able to make adequate justifications for their test in terms of validity and reliability. Since the test design was weak, (e.g. there was no clear model or framework for its development) the validity of the direct test becomes questionable, and all other aspects of the test are affected. One example of this is in the rating criteria/scale. Other than the fact that it was adapted from the MUET scale, lecturers and examiners were uncertain about its purpose and sceptical about its use;

they found the criteria lacking in clarity and not easy to apply. In relation to rater training and standardization, it was found that these were inconsistent across campuses, or none at all in the case of some (see chapter 4, tables4.26 - 4.27). The importance of rater training and standardization as an essential part of a good scoring system is echoed in literature related to speaking (see Weir 1993, 2004/05; Fulcher 2003; Luoma 2004; Brown 2003; Abdul Raof 2003; Dimitrova-Galaczi 2004). The rating conditions under which marking takes place have a potential impact on scoring and need to be standardized too. In the study, problems were related to unfavourable test conditions especially in the branch campuses, which result in inconsistencies such as having only one examiner instead of two, and inappropriate time and place of test. While all examiners and raters were instructed to moderate their marks before finalizing them, this was not possible in cases where only one examiner was present. In terms of grading and awarding, this was not formally recorded in the documentation, and therefore may not be conducted in a similar fashion for every administration. The findings also showed that very little statistical analysis had been conducted on this test, and documents for them were not accessible.

Hence, it is clear that an inadequate set of criteria and rating scale has a direct influence on the raters and the rating process. All other aspects of scoring validity related to the rating procedure have an equal effect on the test, though less directly. Overall, if the conditions for rating are not adequate, systematic and organized, test takers are disadvantaged and test reliability is compromised.

However, it should be noted, that despite the criticism of the original scale, the inter-rater reliability estimate for its use in this project was 0.80, an acceptable level (this should not be taken as an indication of IRR for the test itself, of course). For the computer test, although data was gathered for only one element, i.e. criteria/rating scale, the result for this is positive in relation to the findings, which mainly focused on both the old and new rating scales. This was partly caused by the results (as detailed above) of the direct test findings, which showed that the test was deficient in the other aspects of scoring validity. In essence, because the computer test is new and never before used, the only possible element to explore was in its rateability.
Furthermore, because the crux of the problem with the direct test was related to its criteria/rating scale, a new scale was proposed (see http://www.ets.org for new iBT TOEFL scoring standards for speaking) and its use was explored with members of staff concerned, i.e. lecturers, examiners, experts and administrators (see data/findings from Appendix... Report on Workshop May 2005). Data and findings from the workshop on the rating of the computer test using the new scale produced positive results; staff members were responsive and many were keen to explore the scale further. In fact, the inter-rater reliability estimate for this group of staff members for the computer test was found to be 0.80. This is quite impressive, as the scale used was new to the raters (as opposed to the scale used in the direct test, with which they were very familiar).

Therefore, we can conclude that even though data for scoring validity of the computer test is limited, the potential for its use and further investigation is considerable. As with the results of context validity and theory-based validity, data for scoring validity, though small, is promising. So far, the results of the computer test study had been

positive in terms of context validity and theory-based validity, with more to be investigated for scoring validity. These results point to two central themes of the study.

Firstly, the computer test could be an alternative to the direct speaking test because it is able to address problems arising from the direct test. Secondly, the validation studies on both tests have given us a glimpse of the relationship between the components of validity and their interconnectedness. Most importantly, the computer test had not disadvantaged or affected the students' performance in a negative way; overall correlation between the two tests is 0.70, an indication that a relationship exists between them, in spite of their differences.

Chapter 7: CONCLUSION

7.1 INTRODUCTION

This is the concluding chapter of the thesis on testing of speaking using computer technology in which the findings of the study are evaluated in terms of the research questions, which were formulated as an outcome of the review of literature (Chapter 2). Overall, the study emerged from the position of Mara University of Technology/UiTM) in Malaysia with regard to the existing speaking test which is a direct face-to-face test, and the availability of technology to address problems arising from administering the test to students who are dispersed all across the country. The rationale for exploring the possibilities of delivering the speaking test by means of a computer, though seemingly straightforward, was crucial and related to several concerns. The idea was tied to the university's apprehension with the speaking test which is currently administered as a direct test to thousands of full-time, part-time and distance-learning candidates from all over the country; this presents the university with serious problems in terms of efficiency, reliability and validity. One alternative worth exploring is delivering the test using computer technology and delivering it over the web, in the hope that this could address problems arising from the direct test. In addition (as detailed below in section 7.5), a web-based speaking test could have vast implications for examination boards, universities and the field of language testing within the country, as well as at the international level.

This was the focus of the study, which then encompassed three major areas of investigation: the speaking test (direct and semi-direct), test validation and computerization in testing.

In the past chapters, we looked at the development of the web-based speaking test by examining the validity of an existing direct test using a framework for test validation, which emerged from the work of Weir (2005). Based on the framework, the validation study was conducted, the results were reported according to the validity components, and the computer test was developed, tried, and validated. We presented the findings of the studies conducted on both tests (chapters 4 & 5), and the comparisons were made in chapter 6.

In this chapter, we attended to the significant part of the research, i.e. whether the research questions had been addressed satisfactorily and can be accurately answered. Before such a computer test could be realized, the direct test had to be examined; we needed to investigate it in terms of its validity and reliability, as these are the necessary trademarks for all tests, no matter how small or large scale in nature, and for whatever purpose. The main objective of this investigation was to ascertain if the direct test merits these 'qualities' such that the problems of test context, test administration, security, reliability, and so on could be addressed. It was timely and befitting especially because the test had never been subjected to a formal and systematic validation process. Literature shows that test validation has been standard practice for test developers since the early 50s, but none so far has shown evidence of it in clear and systematic terms, (see Ch 2: Literature Review section on 'Issues of Validity' for details), and so was the case with this university speaking test.

These were concerns shared by the researcher who had been a team member of the speaking test, in all capacity, such as a test developer, administrator and rater. It was thus imperative that a validation study was conducted on the direct test at the beginning of the

research, given the availability of a comprehensive framework for validating language tests, and the speaking test specifically (Weir 2005).

Overall, the discussions in this chapter are made with reference to the research questions. To remind the reader, the research questions are presented below:

Overriding question:

Can an operationalised framework for validating tests of speaking provide an evidential basis for replacing a direct test of speaking ability with a semi direct web based speaking test?

This was then divided into the following questions:

1. To what extent is a face-to-face speaking test valid in terms of:
    a) context validity
    b) theory-based validity
    c) scoring validity

2. To what extent is a proposed semi direct web-based speaking test valid in terms of:
    a) context validity
    b) theory-based validity
    c) scoring validity

3. Can we justify replacing a direct test with a semi direct test of speaking ability?

The remainder of the chapter will present answers to these questions.

7.2 RESEARCH QUESTION 1

In this question, evidence was presented in each of three aspects of validity as suggested in Weir's (2005) framework. Data were gathered from a number of sources using the research instruments, namely the questionnaire, interviews, document analysis, observations, and basic analyses of test scores and rating. All the data are indicated in the table (Figure 7.1) below. (detailed results for the direct test validation study are found in chapter 4)

Figure 7.1 Data gathered for Main Study 1

VALIDITY COMPONENT	Students	Lecturers	Examiners	Administrators	Experts	Documents	Observations
Context validity							
Task Setting	•	•	•	•	•	•	•
Task Demands	•	•	•	•	•	•	•
Task Administration	•	•	•	•	•	•	•
Theory-based validity							
Cognitive Processes	•	•	•	•	•		•
Language Knowledge	•	•	•	•	•		•
Background Knowledge	•	•	•	•	•	•	•
Scoring validity							
Criteria/rating scale		•	•	•	•	•	•
Rating conditions		•	•	•	•		•
Rater Selection		•	•	•	•		
Rater Training		•	•	•	•		
Standardisation/ Accreditation		•	•	•	•		
Rating Decisions (inter-rater agreement)		•	•	•	•		
Consistency (intra-reliability)		•	•	•	•		
Moderation		•	•	•	•		
Grading and Awarding		•	•	•	•		

The results of the data above are summarised below (Figure 7.2) and a discussion follows with respect to each validity component. For each section of the discussion, references are made to issues raised in the current study and the literature on direct testing of speaking.

Figure 7.2 Summary of findings for Main Study 1

Context validity	**Summary of findings**
Task Setting	It appears the test has low validity in terms of most elements of task setting except purpose and order of items
Task Demands	There is lack of validity in terms of most elements of task demands except discourse mode, channel of communication, and interlocutor variables, especially acquaintanceship; however, there is also confusion relating to what these elements really mean for the candidates
Task Administration	The test has clear problems relating to equality in physical testing conditions, uniformity in administration and security across test centres
Theory-based validity	
Cognitive Processes	Students indicated some awareness of processing at planning and speaking time, e.g. able to think of word use, grammatical accuracy, and organization of ideas, and later make adjustments to them while speaking, or monitoring; Staff participants however thought that only very few were capable of these processes and have knowledge of them
Language Knowledge	Students indicated understanding of functional and sociolinguistic knowledge, but not grammatical and discoursal; Staff participants felt they showed this in task B, more than in task A
Background Knowledge	Findings here are divided but more candidates found information from the test useful in fulfilling the tasks; they had minimal recourse to their own background knowledge; staff participants felt the same
Scoring validity	
Criteria/rating scale	The criteria for rating this speaking test are not explicit enough for the examiners to rate the performances efficiently; descriptors are vague and brief, there isn't enough information to help the rater make fair judgements
Rating procedures: Rater Training Standardisation/Accreditation Rating conditions Moderation Statistical analysis	Findings for all elements indicated that the test is significantly lacking in validity for them; rater training and standardization are inconsistent, rating conditions unacceptable in some test centres, moderation is conducted only when there are two examiners present, and minimal analysis is done on the test scores
Grading and Awarding	Validity is low here because of lack of or unavailability of documentation on the process and its outcomes

7.21 DISCUSSION

The issues raised in the literature on the spoken discourse and testing of speaking are interconnected and in many ways, they converge to the same point. That point is related

to the issue of test validity; not only do we measure performance through a test, the test has to be validated in various aspects if we want to ensure that judgements we make about the ability is sound and indisputable. We needed to understand, among other things, processes and resources that learners acquire and utilize to overcome their speech difficulties and to manage conversation as proposed by Brown & Yule (1983), Bygate (1987), Levelt (1989), Tannen (1996), Gass (1997), Cohen (1998), and Hughes (2002), among others. This is crucial so that we can ensure these skills and knowledge are clearly reflected in our tests (see Porter 1991; Weir 1993; McNamara 1997; O'Sullivan 2002; Fulcher 2003; Luoma 2004). Most importantly, this is to ensure that inferences we make of test scores are accurate and point to the candidate's speaking ability and nothing else. In relation to the data and findings above for the direct test, the findings for all components are weak and this is discussed as follows.

♦ **Context validity**

It was clear from all evidence gathered in Main Study 1 for the direct test that the test is weak in terms of its context validity properties, due to several reasons.

First of all, it was discovered that there was no clear model or theoretical framework by which the test was developed; it was adapted from a national test of English (MUET) and minor adjustments were made to suit the university students. The importance of a model/theory-driven test is fundamental to a valid and reliable test. Literature on speaking drives the point that all aspects of a test must be considered if the test is to have any credibility, based on sound theoretical underpinnings and judgement, as proposed by Bachman 1990; Weir 1993, 2005; Bachman & Palmer 1996; Fulcher 2003 and Luoma 2004.

As was discussed in the previous chapters (4 and 6), O'Sullivan (2005) described how the underlying model of ability is operationalised in the test task and rating scale. When this is not apparent, the effect on the test itself and all stakeholders is not constructive, and this was the case with the university speaking test. Lecturers and examiners lacked valuable knowledge needed to enable candidates to maximize their preparations for the test and to provide all candidates with an equal chance of demonstrating their speaking abilities during the test.

This was partly due to inadequate test documentation that was lacking in valuable information for all parties; the test specifications were brief and unclear in its explanations. The literature states that without an explicit set of test specifications, the validity of test tasks remains questionable (see Tarone 1998; Hughes 2002; Fulcher & Reiter 2003; Weir 2005 for in-depth discussions). The test specifications for the university speaking test were lacking in significant aspects of the test and clear instructions for lecturers, examiners and administrators. As a result, our findings showed students and lecturers who were unclear about elements of the test and provided differing information about facets of the test setting such as known criteria for rating, and response format, task demands such as topic/ nature of information, linguistic variables, interlocutor variables, and test administration.

Essentially, the literature on the testing of speaking more recently had focused on the 'oral presentation', advocated as a valuable elicitation task for assessing speaking ability by a number of prominent authorities in the field (Clark and Swinton, 1979; Bygate, 1987;Underhill, 1987; Weir, 1993, 2004; Hughes 1989, 2003; Butler et. al., 2000; Fulcher, 2003; Luoma 2004). Experts have also supported 'interaction' type tasks such as role-play and discussions as being closest to real communication and that they promote language

learning (Long & Porter 1985; Gass & Varonis 1985; Rulon & McCreary 1986; Porter 1986; Long 1989; Hughes 2002; Dimitrova-Galaczi 2004). Various other studies looked at different aspects of interactive tasks such as the effect of interlocutor on the test taker (Porter 1991a, 1991b;Porter & Shen 1991; O'Sullivan 1995; O'Sullivan & Porter 1995; O'Sullivan 2000a, 2000b, 2002), at test taker characteristics and how these affect performance (Berry 1993/94/96; Fulcher 1996; McNamara 1997; O'Sullivan 2002) and how discourse in group interaction is co-constructed between the speakers (Norris & Ortega 2000; Swain 2001; Brown 2003).

Hence, in spite of the abundance of literature on the test of this nature, which contains both the oral presentation and group discussion, it seemed that test developers and experts at the Malaysian university were not equipped with essential knowledge. It was no surprise then that test documentation did not furnish lecturers and students with important information regarding elements of test context; in turn, context validity of the test suffered.

♦ **Theory-based validity**

This aspect of validity includes internal processing (Levelt 1989; Bachman 1990) which the test taker employs when attempting the test task, incorporating his/her own characteristics and background knowledge in the process (O'Sullivan 2000a, for test taker characteristics and how they affect performance). Evidence gathered in Main study 1 from the direct test study showed conflicting testimonies between students and lecturers, examiners and experts. In addition, all participants lacked knowledge in the process of speaking in general, and elements of cognitive processing and executive resource in particular. Like the findings for context validity above, the missing underlying model for

the test caused difficulties for lecturers and examiners, inconsistencies in test administration, and overall, test takers were at a disadvantage.

Essentially, this aspect relates to assumptions that test developers, writers, experts etc. make about students' ability and knowledge of spoken language, what processes they should employ and knowledge they should have (internal, external, linguistic) to be able to perform the tasks well. The underlying model is one that is based on the literature (see Bachman 1990; Weir 2005 for example), and the test designer's task is to establish clearly what operations the candidate is expected to perform and the conditions under which these tasks are to be carried out. Only then, a mismatch will not happen between what is expected of the students in the test (expressed in test task & instructions) and the students' real ability.

The findings above (Figure 7.2) showed that the candidates were rather certain of their strategies and thoughts during the test, but examiners were of a different opinion (not surprising, as one of the examiner's is usually a class lecturer). The complexity of the situation is apparent; the weakness of test design and uncertainties that resulted from this is compounded by the examiners/raters' perception of the candidates and testing situation itself (see McNamara 1997 on the 'social dimension' of interaction in a test, Fulcher 1996 on students' affective responses to different task types). In addition, candidates showed a certain degree of awareness of discourse management during group discussion; for example, they were aware of monitoring their own speech and those of other speakers, the importance of listening to others (audition), and that their responses were, at least in part, influenced by what their peers said. The last point refers to the co-construction of discourse in interaction, which is seen to affect performance and rating (Brown 2003; Swain 2001; Hughes 2002; Fulcher 2003; Luoma 2004; Dimitrova-Galaczi

2004). If this is not made explicit to students during classroom practices and to raters during rater training, effective communication between candidates may not be achieved in the test.

In terms of executive resources, students were not given explicit instructions and knowledge on how to process the task demands, given their own resources in terms of content and language knowledge. While candidates have recourse to their own knowledge when they need to, this knowledge may be limited, in which case, information from the task itself could help candidates process the task. This is not to say that if one side is lacking (internal), the other (external) could compensate; in fact, information on both ends should be appropriate and sufficient for a candidate to fulfil the task. The core of the matter is in the symbiotic relationship between context and theory-based validity (Weir 2005); the context in which test task is presented will influence the internal process of the speaker. Hence, a clear test purpose could affect goal setting and monitoring, response format affects processing, planning time affects goal setting, and even knowing the criteria for rating influences how a candidate approaches the task at hand. Therefore, like the findings for context validity above, the test was ineffective in terms of theory-based validity due to a weak test design, which resulted in test specifications that were not lucid enough for its users regarding cognitive processing for the test. As mentioned in chapter 4, theory-based validity is probably the most difficult to operationalize and investigate, especially since literature on it is also scarce. This however, should not deter test developers from considering it at the test design stage since the test taker is at the heart of the testing process. In the speaking test, the situation is more complex because cognitive processing is in motion while the task is being attempted, and rating takes place at the same time. The complexity of the process can be investigated, for example, as demonstrated by researchers such as Brown (1993), Dornyei & Kormos (1998),

Cohen & Olshtain (1993), and Mackey, Gass & McDonough (2000) among others, who used various introspective methods to elicit data from learners about thought processes involved in carrying out a task.

♦ **Scoring validity**

Like findings for context validity of this test, the scoring validity is also deficient. This leads us back to the link that the validity components in the validation framework have with each other. Essentially, test settings and task demands (context) are perceived by the candidate, who in turn processes all this information, while including his own knowledge of the test content and linguistic demands (theory-based). The elements in test context also have an impact on test criteria (scoring) that we develop and use for rating the test; they should reflect the features of spoken language the test task was designed to generate.

One of the major problems in the university test is in the criteria and rating scale used; it had vague descriptors and lacked content and most of all, how the criteria were developed were not made clear to lecturers and examiners. Weir (2005: 192) asserted that the relationship between task and criteria is essential and they cannot be considered separately at the test design stage. Luoma (2004), Fulcher (1996a, 2003), Abdul Raof (2002), Riggenbach (1998) and Hughes (2002) among others, discussed key issues in developing rating criteria, including the use of professional and expert judgement, and empirical analysis whether qualitative or quantitatively to determine levels of performance within each criterion. A more recent concern has to do with the nature of criteria used for oral assessment, which tend to be biased towards features more easily captured in written performance than spoken (Hughes forthcoming). Hence, test developers need to rethink the criteria/rating scale used for the university speaking test, in terms of its purpose, development and implementation. Related to this is the need to standardize examiners in

their use of these criteria, and this could be incorporated in rater training. In the case of the university speaking test, training and standardization were inconsistent and often non-existent at some test centres (see Brown 2003; Luoma 2004; Rethinasamy 2005, on the importance of standardizing examiners to the criteria used and the rating process). In terms of rating conditions and procedure (see Upsher & Turner 1999; McNamara 2000) the problem was raised in context validity of the university test in terms of uniformity of administration.

Raters' inconsistencies in test administration such as having only one examiner instead of two, allowing more time for preparation, and rating based on student competence rather than performance during the test (as one of the examiner's is usually a class lecturer) , could affect student performance, and rater reliability. Finally, moderation and grading and awarding of marks were also issues which were not addressed efficiently in the university speaking test, thus affecting the reliability of grades (see Weir & Milanovic 2003; O'Sullivan 2006 for details on post-exam grading and awarding of marks).

In conclusion, we are reminded that the crucial issue in the speaking test is what testers want to find out about a student's performance on appropriate spoken interaction tasks. The research is clear that spoken discourse can be assessed in various formats but considerations of test validity and reliability are paramount. In the case of the university speaking test, it was found to be weak in both aspects. The main reason points to the absence of a model or theoretical framework from which the test was designed and developed. As a result, vital information was missing from test documentation, and all participants, directly or indirectly involved in the test, were disadvantaged at different levels and in different ways; because context validity was low, students were not able to demonstrate their best abilities, and rating is affected, which lowers test reliability.

Hence, the test was found to be not valid in terms of context, theory-based and scoring validity.

7.3 RESEARCH QUESTION 2

For this question, evidence was also presented in each of three aspects of validity as suggested in Weir's (2005) framework and data was gathered from a number of sources, namely the questionnaire, discussions, document analysis, observations, and basic analyses of test scores and rating. However, much of the data for this section which were related to the computer test came from students who participated in the computer test trials, pilot study and Main study 2; data from members of staff were gathered at the staff workshop in May 2005 where they saw the test in its entirety, and participated in the rating process and discussions. All the data collected are indicated below (Figure 7.4) (detailed results for the computer test validation study are found in chapter 5)

Figure 7.3 Data gathered for Main Study 2

VALIDITY COMPONENT	Students	Lecturers	Examiners	Administrators	Experts	Documents	Observations
Context validity							
Task Setting	•	•	•	•	•	•	•
Task Demands	•	•	•	•	•	•	•
Task Administration	•	•	•	•	•	•	•
Theory-based validity							
Cognitive Processes	•					•	•
Language Knowledge	•					•	•
Background Knowledge	•					•	•
Scoring validity							
Criteria/rating scale		•	•	•	•	•	•
Rating conditions							•
Rater Selection							
Rater Training							
Standardisation/Accreditation							
Rating Decisions (inter-rater agreement)		•	•	•	•		
Consistency (intra-reliability)		•	•	•	•		
Moderation							
Grading and Awarding							

The results of the data above are summarised below (Figure 7.4) and a discussion follows with respect to each validity component. For each section of the discussion, references are made to issues raised in the current study and the literature on computerized testing.

Figure 7.4 Summary of findings for Main Study 2

Context validity	Summary of findings
Task Setting	Test validity appeared positive in terms of most elements for test setting except weighting (which was not indicated in the test), and time for preparation
Task Demands	Test validity appeared positive for all participants on most of the elements for test task demands
Task Administration	Test validity was positive for physical conditions and uniformity of administration, but not for security
Theory-based validity	
Cognitive Processes	Candidates indicated that, not only did they go through the same processes as they did for the direct test, they were able to process better at some points in the computer test, such as monitoring their word, grammar and organization of ideas while speaking in task A and B. In general, processing was not negatively affected for the computer test, i.e. was not worse than processing in the direct test
Language Knowledge	Candidates were confident of their knowledge in terms of functional and sociolinguistic elements of the test; validity for this aspect appeared positive
Background Knowledge	Candidates' internal knowledge on the topic and information for the tasks were average; they also found information in the test very helpful for test preparation
Scoring validity	
Criteria/rating scale	Raters appeared positive regarding the new proposed scale (TOEFL) 1. Data from staff workshop showed smaller variance for raters marking the direct test using the new scale, compared to the old scale 2. Data from rating both tests (direct and computer) using their own scale showed fairly high rater agreements within a test and between the tests
Rating procedures: Rater Training Standardisation/Accreditation Rating conditions Moderation Statistical analysis	Information available here were on analyses done on data gathered at staff workshop (May 2005) and Main study 2: Correlation index for rater agreement for each test was .08 , and between the tests was .07

7.31 DISCUSSION

It had been mentioned several times in past chapters (1, 3, 5, 6) that although literature on computerized testing is available (such as in Brown 1997; Chapelle 2001; Fulcher 2000; Goodwin-Jones 2000/2001; Norris 2001,; Jones 2001; http://www.llt.msu.edu for issues surrounding computerization in testing), it is not in abundance and usually relates to

more direct forms of language testing. In fact, even fewer studies are found for the computerized speaking test (see Kenyon & Malabonga 1999 -present on CAT for speaking, on-going research projects at UCLA (Bachman et al 2002), DIALANG project (Alderson et al 1998), TOEFL CBT speaking test (TOEFL Academic Speaking Test/TAST)).

However, based on the literature reviewed on the direct, semi-direct and computerized testing (chapter 2), a considerable case was made for the use of a web-based delivered oral proficiency test, such as the one found at the university in Malaysia. Literature has also shown the direct test such as the OPI to have shortcomings in context and scoring validity, and practicality, and semi-direct test such as the SOPI to be weak in context and theory-based validity. Of course, a general concern with the computerized oral test is if it will capture the critical features of speaking performance such that we can make warranted interpretations about the learners' knowledge and ability (Norris 2001). In the case of the present study, a major consideration in its design was that it parallels the original (direct) test as much as possible (see chapter 5, section 5.2 for test design). This is key so that a comparison can be made between them to determine if the computer delivery method would in fact, affect student performance (theory-based), and if other aspects of context validity of the direct test could be enhanced. For scoring the test, we proposed the new iBT TOEFL Scoring standards for speaking (see details in http://www.ets.org) as it has three similar criteria to the old scale and is by far most researched and updated international scale; this is in contrast to the ACTFL Guidelines which has come under a lot of criticism (see chapter2, section 2.31, plus Norris 1997, 2001). In addition, given the scope, reliability and practicality, and validity issues faced by the existing speaking test at the university, an attempt to introduce the indirect test in Malaysia is justifiable, especially because it attends to the concerns of standardization in such large scale testing.

Based on data above, the findings for all components especially for context and theory-based validity appeared positive and better than the results found for the direct test. The discussion that follows is divided according to the three validity components.

♦ **Context validity**

The main advantage of the computer delivered test is in standardization of test input and administration, as summarized above. Data from student questionnaires indicated that candidates responded positively to most elements in test setting, task demands and test administration. They indicated that purpose, instructions and information for the tasks were clear, and variables relating to the interlocutors in task B did not hamper their responses. Most of all, students were pleased with the test conditions and administration, all of which were major factors raised in the direct test administration. In the computer test, students' anxiety and nervousness were reduced with the absence of the examiner (s) (see Malabonga, Kenyon & Carpenter 2005, on how COPI presents a more flexible and less threatening experience for examinees).

Staff interview data indicated that they recognized the standardization factor in the computer test, which is an advantage in itself for several reasons. Firstly, it removes the burden of producing six equivalent forms of question papers each year for the direct test (see O'Sullivan, Weir & Horai 2004 on difficulty in establishing task equivalence). Secondly, because all candidates received the same input in terms of topic, information in tasks, linguistic variables and interlocutor variables, variability resulting from task differences and uniformity of administration is removed, and interlocutor training becomes unnecessary. Because candidates do not interact with a live interlocutor, co-construction of discourse does not take place in the computer test. Literature is clear on

how this causes a major problem for rating in interactive type tasks as discussed in Lumley & Brown (1997), Brown & Hill (1998), Brown (2003), Swain (2001), Fulcher (2003), Luoma (2004), and Dimitrova-Galaczi (2004); however, not a lot has been said about how this can be addressed.

The analysis of the existing test performances indicated that there was little to be found in terms of interactive functions; this indicated that one of the primary reasons for including so called 'interactive' tasks in the computer test was redundant. Thus, lack of interactive language in the direct test made replicating the original test on the computer easier. There is also a permanent record of students' performance, which could be re-evaluated if needed and used for future rater training. Most of all, for this speaking test, which is administered to twelve campuses across the country, standardization is possibly its leading advantage.

Hence, it appears at this point that data indicates high context validity for the computer test. Mainly, the computer test is able to address the problems raised in the direct test relating to test unfairness due to lack of equivalence in test forms, test reliability due to problems with co-construction of discourse and a weak rating scale, inconsistent test administration, and task demands which were unclear and often unpredictable for the candidates.

♦ **Theory-based validity**

In general, our findings indicated that not only did students engage in internal processing in the same way that they did for the direct test, they were also able to perform certain strategies better in the computer test, especially for task B. One of the main reasons seemed to be that the absence of examiner/interlocutor made monitoring and articulating

less stressful as the candidates felt less intimidated (Malabonga, Kenyon & Carpenter 2005; Shermis & Lombard 1997; Henning 1991). The same results were found for task B, which for these candidates would have been the first time being presented with such a format for an 'interactive' task. Hence, one of the main features of the computer test, its ability to replicate the direct test while controlling for task demand and setting, is an advantage to our candidates.

Literature on computerized testing, though again scarce for the speaking test (except for on-going work on the COPI, Kenyon et al 1999- present, and iBT TOEFL Speaking test/TAST), constantly reminds us of the main consideration of construct validity, i.e. whether the computer can capture the underlying construct of speaking. Weir (2005) asserted that construct validity is central and a test is valid if the elements of the validity components (context, theory-based, scoring) are evident and proven effective and purposeful for the test. Most other authorities on computerized testing have echoed the same concern though in terms of future computer-based projects (Fulcher 2000; Norris 2001; Roever 2001; Dunkel 1999). Since the computer test used in this research is a prototype, and was used for the first time at the university in Malaysia, the findings showed that students had not been seriously affected in terms of their internal processing and knowledge of the tasks.

It was also noted (chapter 5) that evidence is overwhelming that candidates go through some form of cognitive processing as described in Levelt (1989). This process is not linear and straightforward but in fact involves utilizing several elements at the same time. Monitoring is important as speakers monitor their own words to make sure they do not make errors. Speakers also make spontaneous self-repairs as they attend to various aspects of the action they are performing (see Field 2003); they monitor the

message/concept, the way it should be expressed to avoid ambiguity (discoursal), level of formality required by the context of discourse (sociolinguistic), lexical, syntax and morphological (grammatical). Our candidates indicated that they went through the conceptualization phase during preparation time for the task while monitoring its content and linguistic demands, and they monitor and self-repair their own language use and content knowledge before as well as at the point of articulation.

Therefore, it is clear at this point that candidates go though internal processing when attempting the test tasks and that most processes were similar to the ones they employed during the direct test. In addition, the computer test has the advantage of removing the interlocutor factor, which makes self-repair easier and less stressful for the candidates, and it is able to replicate the direct test while controlling the effects that test setting and task demands might have on the candidates.

♦ **Scoring validity**

Two elements were collected under scoring validity for the computer test: the criteria/scale for rating and basic analysis on its application. One of the major issues that emerged in the direct test validation study was rating; the test was deficient in most aspects of criteria/rating scale, rater concerns, and rating conditions and process. Literature on developing rating scales and rating emphasizes the importance of empirical research and validation (Stansfield & Kenyon 1995; Henning 1996; Fulcher 1996a; Halleck 1996; McNamara & Adams 1996; North & Schneider 1998). We need to consider among others, criteria that reflect processing conditions that the task requires (Luoma 2004; Weir 2005), levels of proficiency that student performance will result in (see ETS 2000 TSE scale; Weir 1993; 'Cambridge Common Scale for Speaking' in Cambridge ESOL Main Suite exam handbooks), and in interaction task, how rating is invariably given to individuals

when discourse is co-constructed between participants (Lumley & Brown 1997; Brown & Hill 1998; Brown 2003). The importance of ensuring that test criteria reflect the demands of test tasks is again, reiterating the relationship between context validity and scoring validity (Lynch & McNamara 1998; Chalhoub-Deville 1995; Upsher & Turner 1995; Fulcher 2003; Luoma 2004; Weir 2005).

Because of major problems with the old scale, it was proposed that a new scale be used as a rating tool for the computer test. The new iBT TOEFL Scoring standards (ETS 2004) was selected as a starting point for developing a more balanced scale for UiTM, mainly because it has three similar criteria to the old scale, is well-researched, and used by a major examination board. Preliminary data on the use of the TOEFL scale by university staff members showed encouraging results (see Appendix 3.10). It appeared that in spite of the fact that participants were new to the scale and had no prior training, they were able to use the scale effectively. Further ratings of the computer test on two occasions (staff workshop and Main Study2), using the new scale produced reliability indexes which were acceptably high. While these results may point to the rating for this group of participants in this study, they are nevertheless positive, and the use of a new scale must be considered.

Thus, it is fair to say that response from participants regarding the scoring validity of the computer test was positive. In fact, participants were able to realize the potentials of the computer test where scoring validity is concerned such as:

- The reliability of scoring could increase, for e.g. pass/fail cases can be easily identified using the scale
- Rating is conducted later so inter-rater reliability problem is removed, and there is a permanent record so tests can be re-played & re-evaluated

- The potential for standardization of marking across the board, especially across campuses is possible

In conclusion, it is imperative that those involved in test development are absolutely clear about the performance they want to find out of the students, regardless of the mode of delivery. Research on computerized testing is limited but clear in that testing spoken discourse using computer or web-based technology is the way forward. However, as our study above advocates, a systematic validation process is needed before claims can be made of test validity or reliability. While our findings showed positive results for context, theory-based and a small part of scoring validity of the computer test, we also recognized from the direct test study that the absence of an ability model can make a test atheoretical. The computer test had the advantage of a set of specifications that were drawn up based on revised specifications for EXAVER levels 1-3, University of Veracruz (EXAVER 2005), including one for test taker's cognitive processing and background knowledge, and one for scoring (see Appendix 3.12). All these documents were prepared during the test development stage, and were based on the framework for test validation. In this way, and in proposing a new scale which had been professionally researched and developed, and is based on essentially the same construct as the tasks, the computer test stands on a firm position and with added advantage.

Therefore, we conclude that the computer-delivered speaking test was valid in terms of context and theory-based validity, and for at least a small part of scoring validity.

7.4 FINAL QUESTION/ DISCUSSION

With reference to the third question, we can safely state that it is justifiable to replace the direct test with a semi direct, web-based test of speaking ability. Our findings from the two major studies (direct test and computer test validation) illuminated the on-going problems faced by the direct test, which the computer test attempted to resolve.
In essence, we can now respond to the overriding question in the study, see Section 7.1 above.

It was evident throughout the study that the framework for validating speaking as suggested by Weir (2005) was useful in terms of its purpose and applicability to the study. We were able to operationalize its content into research instruments for the study, which then enabled us to gather evidence as a basis for replacing the direct test of speaking with a semi direct web-based speaking test. We were able to discover pressing issues that a large-scale speaking test such as the Malaysian test face, systematically and objectively. Beyond this, we showed that the web-based speaking test is a potential alternative and it has the major advantage of addressing the concerns of standardization in large scale testing.

Therefore, we feel strongly, that the validation framework is valuable and applicable for its purpose, and the web-based speaking test is comparable with the direct test. Consequently, based on the availability of a test validation framework and the considerable findings of the current study, we have realized the significant contributions that computerization can make on testing spoken language.

Firstly, the growing literature on the features and potentials of computerization in testing and assessment supports this. For example, experts at CAL (Kenyon et al 1999- present) have developed a CAT for speaking, and contributors to *Language Learning & Technology* constantly discuss issues related to tests that are computer assisted or adapted (see Brown 1997; Dunkel 1999; Chalhoub-Deville 2001; Goodwin-Jones 1997- present; Roever 2001; Norris 2001; among others). Furthermore, literature on comparability of conventional tests (usually paper-based language test/PBLT) and computerized test (usually computer-based language test CBLT) have shown that the differences are usually insignificant, and/or the tests measure the same construct (see Choi et al 2003; McDonald 2001; Russell 1999; Russell & Haney 1997/2000; Neuman & Baydoun 1998; Mead & Drasgow 1993; among others). Moreover, research has also shown that test takers have a general anxiety towards the testing process itself, rather than what's termed 'computer anxiety', or anxiety due to other factors such as age, gender and personality (see Thelwall 2000; Shermis & Lombard 1997; Paul 1994; among others).

In terms of scores obtained by the candidates who participated in the computer speaking test, the data showed overall correlation between the tests (direct and web-based) to be acceptably high (0.70). It gives an indication of the comparability between the two tests, and the rating scales that were used. On the whole, these results demonstrate the fact that the web-based speaking test is comparable to, or even better than the direct test in terms of its construct validity.

7.5 IMPLICATIONS OF THE STUDY

This study on web-based testing involved developing a framework for validating the existing direct speaking test, and more importantly, developing a semi-direct web-based test which is as good, if not better, than the direct test in all aspects of the framework.

Unfortunately, not only is the literature on computer-based language testing limited, the literature on validation of this mode of testing is almost non existent (though, see Kenyon & Malabonga, 1999, 2000, 2001, 2002). The literature on the vital issue of test validity has been more comprehensive (Messick 1975, 1980, 1981, 1988, 1989, 1995, 1996; Bachman 1990, Bachman & Palmer 1996; Brown 1996; Grabe & Kaplan 1996; Guerrero 2000, Hasselgren 2000, 2004; Weir 2005), but we still lack a comprehensive and operationalisable framework that can be used to validate the speaking test conducted at tertiary level with ESL speakers. This thesis adapted and empirically trialled a new theoretical framework for exploring the validity of such speaking tests. This involved innovative work in exploring the socio-cognitive aspects of speaking in English, and developing a new model for validating what a speaking test in English involves. As such, it represented a major advance over traditional, largely cognitive approaches to the testing of speaking. It incorporated for the first time, critical contextual parameters that are shown to have an effect on student language performance in speaking, as well as integrating reliability and consequential validity considerations into the new model.

A major problem occurs in the context of this study, namely the procedures, administration, and logistics of a large-scale direct speaking test. The introduction of an innovatory web-based speaking test, could address this problem; more students can be reached without them having to make special arrangements to do the test.
The outcome of this study is also of interest to governing bodies in the country. It is of particular interest to the Department of Education, who now offers the Malaysian University English Test/ MUET test to all those who wish to pursue higher education in any of the local universities in Malaysia. The MUET test has a speaking component, and a web-based speaking test would enable the department to reach out to many more candidates throughout the country. They would not have the massive task of organizing

and administering the test to more than twenty thousand students at a time, twice a year, at various centres, any more.

The nation has been advancing aggressively toward the electronic mode via the web in major sectors of education, health, information transfer, telecommunication, just to mention a few (see, for example Economic Planning Unit 2006). This is evident in the government's 'smart schools' project nationwide which was considered as top priority in the government's colossal MSC (Multimedia Super Corridor) venture of the early 1990s. The introduction of a test which is computer driven can benefit all parties involved, especially educators, learners, and administrators in these government sectors. In the private sector, organizations involved in distance-learning education, such as the Tun Abdul Razak University, would benefit from successful web-based testing as they would be able to conduct tests to their candidates who come from different parts of the country.

If the delivery of tests of speaking on-line proves a success, examination boards may be able to use the same method for testing spoken language, and like UiTM and the Department of Education, will be able to cater for many more candidates using this method. Most importantly, the study will also provide important empirical evidence on the validity of the speaking test that is conducted at UiTM, Malaysia. Although it is only a component of the proficiency English course offered by the language centre, it is a major component, and has been given very little attention in terms of validation. In fact, the distance learning centre which has students at all the branch campuses in the country has been in existence for nearly ten years, and no validation has been carried out on the direct speaking test for these students. The findings of this study could contribute to the centre's on-going process and efforts in research and development to improve on the quality of its

services to the customers. The more global implications of the findings of this work are presented in section 7.7 below.

7.6 LIMITATIONS OF THE STUDY

The breadth of the study in terms of the major areas which it covered: testing of spoken language, test validation and computerized testing, is the strength and weakness of the study.

While its strength is in the originality of the design and methodology aspects which involved the application of a new validation framework to make comparisons between two tests, it was also a weakness. The validation framework is the first such framework which considers in detail what the test taker brings to the test, a socio-cognitive approach, and each validity component is described in detail in terms of the elements they each entail, which makes it comprehensive.Unfortunately, these aspects of the framework require a great deal of time and effort on the part of the researcher to materialize in the study. The number of elements in the framework and the relationship between them are conceptual considerations, and the application of these elements into research instruments is a complex, practical task.

Much of the data gathered in the validation studies were qualitative in nature, more than quantitative. Again, due to its coverage and broad perspectives taken in terms of aspects of validity of the two tests, we were restricted in the range of research methods that we could employ. For example, data from the direct test indicated that its scoring validity was weak, hence, we did not feel the necessity to analyse test scores and rater reliability further, as the rating scale used and scoring procedures were faulty in many ways.

In essence, it may have been able to employ more research techniques and a more in-depth use of the validation framework given more space and resources, especially time; these were not achievable within the space of the dissertation.

However, the potential of the computer test using web-based delivery is one of the major contributions of the study. It was apparent from the findings of the validation study on the direct test that the two tasks (oral presentation and group discussion) were different constructs, and more work is needed to identify these differences, especially since at least conceptually they would require separate rating scales and attention to rater training.

On reflection, in terms of limitations of the study, the two most prominent were related to resources and to the scope of the study.

7.6.1 Resources

One of the constraints here is that of time. As the researcher collected most of her data at the university in Malaysia, travelling twice a year out of the UK was time consuming, and the times for collecting data had to coincide with the times when the direct test was conducted each year, i.e. March and September. More importantly, the researcher was working in collaboration with colleagues (lecturers/examiners/administrators, etc) who gave up their time and students for the study; this was not possible all the times. Related to this is the scale of the study which was big as it encompassed research into test validation, the speaking test, and computerized testing. More time and resources would be needed to investigate each area thoroughly, which could not be achieved in the limited time awarded for completing the study, and in the space of a dissertation. However, in spite of these constraints, a substantial amount of data was gathered as the research was

approached and organized systematically; relevant parties such as the director of the Language Centre and the registrar's office at the university were officially informed of data gathering. Cooperation and support from lecturers, administrators and students were beyond expectations, and this contributed valuable findings to the study.

7.6.2 Scope

This is related to the framework used for validating the speaking test; it was applicable as a tool but its practicality was a limitation. The framework consists of five components, each with its own elements. For the research, three components were used (context, theory-based and scoring validity) as these are seen to reflect construct validity. Each one had between 10 - 22 elements; these were incorporated into the research instruments, especially the questionnaires. As a result, the amount of data collected for the whole study was considerable, in spite of the fact that there were two such validation studies including trials and pilot studies; hence, it was not possible to allocate sufficient time focussing on any one component or element in particular. However, despite this limitation, we were able to obtain a broad perspective of the issues relating to testing of spoken language, which was valuable to the conclusions of the study, and a detailed account of the direct and web-based tests were in fact obtained according to the validity elements of the framework.

7.7 CONTRIBUTIONS TO THE LITERATURE

The international significance of this study resides in its broad-ranging approach. First, the development of a validation framework for speaking is valuable for evaluating all international high stakes tests of overseas students' language ability, specifically, their speaking ability, e.g. IELTS, TOEFL and CPE, that are used worldwide for university matriculation purposes. It will set the criteria against which all such tests might be judged

and provide the methodology for carrying out such studies. At the moment, these tests are largely accepted on trust and not subject to the rigorous scrutiny that a validation framework can provide. In this study, we were able to operationalize the framework in terms of the research instruments used to gather all its data, and consequently, suggest a methodology for making comparisons between tests.

Second, the development of a computer delivered speaking test using web-based technology is a major contribution to the field of language testing, especially for the international community of researchers, test developers, and major testing bodies such as ETS and Cambridge ESOL. It would contribute to the field of web-based testing by establishing whether a speaking test in this mode of delivery can perform as well, if not better, than the direct test in the specified aspects of the theoretical framework for speaking. If this new procedure results in a test that is less anxiety-ridden, and can clearly 'test for best' (Swain 1986), then it can be of real benefit to all test takers.

Finally, in terms of the current project the ultimate beneficiaries of such a test are the test takers in distance learning programmes who will have enhanced geographical access to the new procedure. Essentially then the thesis will contribute to the creation of a set of standards for the development, administration and use of tests of spoken language in English. These do not exist at the moment and we suffer from the consequences of this.

7.8 FUTURE RESEARCH

As much as we were able to obtain information for the study such that we could formulate valuable findings and conclusions from them regarding test validation, validating the speaking test, and computerized testing, there are areas which would benefit from further exploration. Mainly, these are related to key findings of the direct and

indirect, web-based speaking test. They could be expressed in terms of the following propositions.

7.8.1 *The use of the validation framework*

Although the framework proposed by Weir (2005) was comprehensive and detailed in its illustration of the validity components, its significance could not be fully realized in a study such as the current project, which is limited in its time and space. Thus, future research on test validation could be conducted by incorporating all the validity components or by employing one component at a time; in each case, more time and resources are to be expected. There were issues here with the attempt to operationalise the various parameters within the framework. These issues suggest that the parameters have yet to be fully understood.

7.8.2 *The consequence of a theoretical model which drives the test*

We discovered that the UiTM speaking test was deficit in its construct validity mainly due to the absence of a clear model by which the test was designed. A new model (O'Sullivan 2005) was proposed, which illustrates how the underlying model or theory of ability which steers the test is directly linked to test tasks and the rating scale, which themselves affect rating, performance and the eventual meaning that we make of the test scores. Further research on the use of this model for test design and development could benefit language testers, developers, experts, etc for future development of a more reliable and valid speaking test.

7.8.3 *The computer or web-based test of speaking*

The computer-delivered speaking test developed for the project, though at its infancy stage, was prototypical of the capabilities that such a test has over the direct method of testing speaking. It has also opened avenues for further research that is needed. One area is in establishing a comprehensive platform for delivery, which could incorporate audio, video, text, pictures, all together or in various combinations in the test. In addition, we need a platform that has the ability to store and disseminate recorded performances reliably and efficiently, and is secured.

In terms of test tasks, the study highlighted problems of interactive tasks in direct speaking tests. Mainly, the problems relate to co-construction of discourse, which affects rating, and the lack of or absence of interactional speech functions and agenda management, which are expected outcomes of interaction. Hence, a task was built into the computer test such that even though candidates did not take part in live interaction, they were expected to recognize the functions as they appeared in the test. Until further research into computer capabilities for 'live' interaction is conducted, we could investigate the features of speech in interaction and how to capture these in the web-based test.

7.8.4 *The criteria/rating scale for the speaking test*

This was one of the major findings of the study, and even though the new TOEFL scale was proposed for both the direct and semi-direct computer tests, it was not developed for either of the tests; it was however, argued that the scale was suitable due to the similarity of language model and construct that supports the scale. What is needed is an investigation into the development of a new scale, especially for the computer test, which takes into account fundamental considerations such as those proposed in Weir (2005: 178-179; 191-199).

7.8 CONCLUDING REMARKS

One of the biggest contributions of this study is in the wide implications that the research on developing and validating the web-based speaking test could have on future research, examination boards, as well as language teaching and testing as a whole. It had an innovative approach in its design, development and methodological considerations.

Firstly, the theoretical basis for its design and development were based on literature in the areas of computer -based and web-based testing; these were reviewed and their potentials for overcoming the problems arising from the direct test were explored. In addition, a set of specifications were developed for the test in terms of its contextual features, cognitive processing a successful candidate is expected to employ in attempting the tasks, and scoring considerations for the test. Moreover, an alternative rating scale was proposed for use in scoring the test as it approximated the test in terms of language model and construct. For validation it had the advantage of a socio-cognitive framework for validating a speaking test, which takes into consideration key elements of construct validity and proposes a systematic and organized method for establishing test validity.

Therefore, the web-based test was designed and developed on a sound theoretical background. As described in chapter 3, it was administered with adequate support and infrastructure, and data were analysed according to the elements of construct validity. More importantly, while this is a first attempt at such a method of testing general spoken language ability, it offers an alternative to the traditional method without jeopardizing the validity and reliability of the speaking test.

Another major outcome of this study was in its ability to operationalize the framework for test validation. This process involved refining the framework and its validity components

in preparation for developing research instruments which reflected these components; this was crucial if we wanted to ascertain the validity of a test more comprehensively. This was a first attempt at translating the framework into research instruments, such as the speaking test questionnaire, and though the process was tedious and time-consuming, the outcome was significant to the study. The amount of data obtained was substantial and provided us with a broad spectrum of the tests (both direct and web-based) and their validity properties.

While a lot more work is needed in these areas, the outcomes of this research have paved the way for further investigations into:

- the advantageous features of the computer speaking test using web-based technology, and
- how to systematically validate the test such that the critical features of speaking are captured, ensuring that inferences we make regarding student performance are reliable and justifiable.

By undertaking such investigations as were undertaken in this study, we were able to make substantiated claims about the direct speaking test, and more importantly this enabled us to develop a semi-direct web-based test which was innovative and has far-reaching consequences in language teaching and testing.

BIBLIOGRAPHY

Abdul Raof, A. H. (2002). The Production of a Performance Rating Scale: An Alternative Methodology. School of Linguistics and Applied Language Studies. Reading, University of Reading

Abraham, R. G. and H. Liou (1991). Interaction generated by three computer programs: Analysis of functions of spoken language. New York, Newbury House.

Adams, M. L. (1978). Measuring foreign language speaking proficiency: A study of agreement among raters. Direct testing of speaking proficiency: Theory and application. J. L. D. Clark. Princeton, NJ, Educational Testing Service: 131-149.

Adams, M. L. (1980). Five Cooccuring Factors in Speaking Proficiency, Georgetown University Press.

Aguirre-Munoz, Z. & Baker, E.L. (1997). Improving the Equity and Validity of Assessment-Based Information Systems. Challenges Minorities face in Educational Testing and Assessment. M. Nettles. Boston, Kluwer.

Alderson, J. C., Ed. (1993). Judgements in Language Testing. A New Decade of Language Testing Research, TESOL.

Alderson, J. C. (2000). Assessing Reading. Cambridge, Cambridge University Press.

Alderson, J. C. et al. (1998). DIALANG, The European Commission.

Alderson, J. C., Clapham, C. & Wall, D. (1995). Language test construction and evaluation. Cambridge, CUP.

Alderson, J. C., Krahnke, K. J. and Stansfield, C. W. (1987). Reviews of English language proficiency tests Washington, D.C. , Teachers of English to Speakers of Other Languages.

Alderson, J. C. & Buck, G. (1993). "Standards in testing: A study of the practice of UK examination boards in EFL/ESL testing." Language Testing **10**(2): 1-26.

Alderson, J. C. & Hamp-Lyons, L. (1996). "TOEFL preparation courses: A study of washback." Language Testing **13**(3): 280-297.

Anastasi, A. (1988). Psychological Testing. New York, Macmillan.

Bachman, L. (1973). "Testing oral production." Bulletin of the English Language Centre (Bangkok) **3**(1): 41-58.

Bachman, L. (1990). Fundamental considerations in language testing. Oxford, OUP.

Bachman, L. & Cohen, AD. (1998). Interfaces between second language acquisition and language testing. Cambridge, CUP.

Bachman, L. et al. (1995). An investigation into the comparability of two tests of English as a foreign language. Cambridge, CUP.

Bachman, L. F. (2002). "Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments." Educational Measurement: Issues and Practice: 5-18.

Bachman, L. F. & Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. The construct validation of tests of communicative competence. P. J. M. G. A. S. Palmer, & G. A. Trosper. Washington DC, TESOL: 149 - 165.

Bachman, L. F. & Savignon, S. (1986). "The evaluation of communicative language proficiency: A critique of the ACTFL oral interview." Modern Language Journal **70**: 380 - 390.

Bachman, L. P., AS (1996). Language testing in practice. Oxford, OUP.

Bachman, L., Carr, N., Pan, M., Vongpumivitch, V., and Xi, X. (2002). Validity issues in a web-based language assessment system (WebLAS). 24th Language Testing Research Colloquium, Hong Kong.

Bailey, K. M. (1996). "Working for washback: A review of the washback concept in language testing." Language Testing **13**(3): 257-279.

Baker, R. (1997). "Classical test theory and item-response theory in test analysis." Language testing update.

Ball, F. W., J. (2002). "Research projects relating to YLE Speaking tests." UCLES EFL Research Notes **7**: 8-10.

Barnwell, D. (1989). "Proficiency and the Native Speaker." ADFL Bulletin 20(2): 42-46

Bazen, D. (1979). "The place of conversation tests in oral examinations." English in Education **12**(2): 39-50.

Beardsmore, H. B. (1974). "Testing oral fluency." IRAL **12**(4): 317-326.

Beattie, G. W. (1981). "Interruption in conversational interaction, and its relation to the sex and status of interactants." Linguistics **19**: 15 - 35.

Bell, J. (1987). Doing your research project. Milton Keynes, Open University Press.

Berry, V., Ed. (1993). Personality characteristics as a potential source of language test bias. Language Testing: New Openings. Jyvaskyla, Finland, Institute for Educational Research.

Berry, V. (1994). Personality characteristics and the assessment of spoken language in an academic context. Language Testing Research Colloquium.

Berry, V. (1995). A qualitative analysis of factors affecting learners' performance in group oral tests. Language Testing Research Colloquium.

Berry, V. (1996). Ethical considerations when assessing oral proficiency in pairs. LTRC 96, Japan.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). "Speaking and Writing in the University: A Multidimensional Comparison." TESOL Quarterly **36**(1): 9-47.

Bonk, W. J. O., G.J. (2003). "A many facet Rasch analysis of the second language group oral discussion task." Language Testing **20**(1): 89-110.

Booth, D. (2002). "Revising the Business English Certificates (BEC) speaking tests." UCLES EFL Research Notes **8**: 4-7.

Booth, W. C., Colomb, G. C. & Williams, J. M. (2003). The Craft of Research. Chicago, The University of Chicago Press.

Borsboom, D. & Mellenbergh, G. J. (2004). "The concept of validity." Psychological Review **111**(4): 1061 - 1071.

Bortfield, H., Leon, S. D., Bloom, J. E., Schober, M. F. & Brennan, S. E. (2001). "Disfluency rates in conversation: Effects of age, relationship, topic, role and gender." Language and Speech 44(2): 123 - 147.

Bradin, C. (1999). "Review of 'Oral Language Archive'." Language Learning & Technology **2**(2): 16 - 22.

Brennan, S. E. et al. (2001). "Disfluency rates in conversation: Effects of age, relationship, topic, role and gender." Language & Speech **44**(2): 123-147.

Brooks, L. (2003). "Converting an Observation Checklist for use with the IELTS Speaking test." UCLES EFL Research Notes **11**: 20-21.

Brown, A. (1993). "The role of test taker feedback in the test development process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese " Language Testing **10**: 277 - 304.

Brown, A. (2003). "Interviewer variation and the co-construction of speaking proficiency." Language Testing **20**(1): 1-25.

Brown, A. (2004). Interviewer variability in oral proficiency interviews, The University of Melbourne, Melbourne.

Brown, A., & Lumley, T. (1996). "Interviewer variability in specific-purpose language performance tests." Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96: 137 - 150.

Brown, A. & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. Sydney, ELICOS: 173-91.

Brown, C. M., & Hagoort, P. (1999). The neurocognition of language. Oxford, Oxford University Press.

Brown, G., & Yule, G. (1983). Teaching the Spoken Language. Cambridge, Cambridge University Press.

Brown, G. & Yule, G. (1994). Assessing spoken language. Teaching the Spoken Language. Cambridge, Cambridge University Press.

Brown, J. D. (1988). Understanding research in second language learning. Cambridge, UK, CUP.

Brown, J. D. (1996). Testing in language programmes. New Jersey, Prentice Hall Regents.

Brown, J. D. (1997). "Computers in language testing:Present research and some future directions." Language Learning & Technology 1(1): 44-59.

Brown, J. D. (2001). Using Surveys in Language Programs. Cambridge, Cambridge University Press.

Brown, J. D. & Rodgers, T.S. (2002). Doing Second Language Research. Oxford, Oxford University Press.

Buck, G. (2001). Assessing Listening. Cambridge, Cambridge University Press.

Butler, A., Eignor, D., Jones, S., McNamara, T. & Suomi, B. (2000). TOEFL 2000 Speaking Framework, ETS.

Bygate, M. (1987). Speaking. Oxford, Oxford University Press.

Bygate, M. (1999). "Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks." Language Teaching Research 3(3): 185-214.

Bygate, M. (2003). Oral Language Content and Oral Language Learning: Issues in the Teaching of Spoken Language 8th International Bilkent University School of English Language ELT Conference, Turkey, Bilkent University.

Byrnes, H. (1991). Recent trends in language testing: The case of testing oral language, Language Laboratory Association of Japan.

Byrnes, H. & Canale, M. (1987). Defining and developing proficiency: Guidelines, implementations and concepts. Skokie, National Textbook Company, IL.

Canale, M. (1988). "The Measurement of Communicative Competence." Annual Review of Applied Linguistics 8: 67 - 84.

Canale, M. & Swain, M. (1980). "Theoretical Bases of Communicative Approaches to Second Language Teaching and testing " Applied Linguistics 1(1): 1 - 47.

Carroll, J. M., Tanenhaus, M.K., & Bever, T.G., Ed. (1978). The perception of relations: The interaction of structural, functional, and contextual factors in the segmentation of sentences. Studies in the Perception of Language. New York, Wiley.

Carter, R. & McCarthy, M. (1995). "Grammar and the spoken language " Applied Linguistics 16 (2): 141-158.

Carter, R. & McCarthy, M. (1997). Exploring Spoken Language. Cambridge, Cambridge University Press.

Carter, R., Hughes, R. & McCarthy, M.J. (2000). Exploring Grammar in Context Cambridge, Cambridge University Press.

Cascallar, M. I. (1996). Modified oral proficiency interview: Its purpose, development and description. LTRC 96, Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96.

Chalhoub-Deville, M. (1995). "A contextualized approach to describing oral language proficiency." Language Learning **45**(2): 251-281.

Chalhoub-Deville, M. (2001). "Language testing and technology: Past and future." Language Learning & Technology **5**(2): 95-98.

Chalhoub-Deville, M. (2003). "Second language interaction: Current perspectives and future trends." Language Testing **20**(4): 369-383.

Chalhoub-Deville, M. & Deville, C. (1999). "Computer-adaptive testing in second language contexts." Annual Review of Applied Linguistics **19**: 273 - 299.

Chambers, F. & Richards, B. (1996). "Reliability and validity in the GCSE oral examination." Language Learning Journal **14**: 28-34.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. Interfaces between second language acquisition and language testing research. L. C. Bachman, AD. New York, CUP: 32-70.

Chapelle, C. (1999). "Validity in language assessment." Annual Review of Applied Linguistics **19**: 254-272.

Chapelle, C. (2001). Computer applications in second language acquisition. Foundations for teaching, testing and research. Cambridge, UK, CUP.

Chapelle, C. & Douglas, D., Ed. (1993). Foundations and directions for a new decade of language testing. A New Decade of Language Testing Research, TESOL.

Chapelle, C. A., Enright, M. K. & Jamieson, J. (2004). Issues in Developing a TOEFL Validity Argument. Princeton, Educational Testing Service: 1 - 18.

Chapelle, C., Jamieson, J., & Hegelheimer, V. (2003). "Validation of a web-based ESL test." Language Testing **20**(4): 409-439.

Chaudhary, S. (1997). "Testing spoken English as a second language." Forum **35**(2): 22-29.

Cheepen, C. & Monaghan, J. (1990). Spoken English: A Practical Guide. London, Pinter Publishers.

Cheng, L. (2001). Bringing about changes in language teaching through changes in language testing, Queen's University Canada.

Choi, I., Kim, K.S. & Boo, J. (2003). "Comparability of a paper-based language test and a computer-based language test." Language Testing **20**(3): 295-320.

Chomsky, N. (1965). Aspects of the theory of syntax Cambridge, Mass., MIT Press.

Chomsky, N., Ed. (1976). On the biological basis of language capacities. The Neuropsychology of Language. New York, Plenum Press.

Chomsky, N. (1980). Rules and Representations. Oxford, Basil Blackwell.

Clark, J. L. D. (1978). Interview testing research in Educational Testing Service. Princeton, Educational Testing Service.

Clark, J. L. D., Ed. (1979). Direct vs. semi-direct tests of speaking ability. Concepts in language testing: Some recent studies Washington, DC, TESOL.

Clark, J. L. D., Ed. (1980). Towards a common measure of speaking proficiency. Measuring Spoken Language Proficiency. Washington DC, Georgetown University.

Clark, J. L. D. (1988). "Validation of a tape-mediated ACTFL/ILR scale based test of Chinese speaking proficiency." Language Testing **5**: 187 - 205.

Clark, J. L. D., & Clifford, R. T. (1988). "The FSI/ILR/ACTFL Proficiency scales and testing techniques: Development, current status, and needed research." Studies in Second Language Acquisition 10: 129-47.

Clark, J. L. D., & Li, Y.C. (1986). "Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages. Washington, DC, Center for Applied Linguistics.

Clark, J. L. D., & Swinton, S.S. (1979). Exploration of speaking proficiency measures in the TOEFL context. Princeton, NJ, Educational Testing Service.

Clark, J. L. D., & Swinton, S. (1980). The Test of Spoken English as a measure of communicative ability in English-medium instructional settings. Princeton, NJ, Educational Testing Service.

Cohen, A. D. (1994). Assessing language ability in the classroom. Boston, Heinle & Heinle.

Cohen, A. D. (1996). "Towards enhancing verbal reports as a source of insights on test-taking strategies." Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96: 339 - 365.

Cohen, A. D. (1998). Strategies in Learning and Using a Second Language. London, Pearson Education Limited.

Cohen, A. D. & Olshtain, E. (1993). "The Productions of Speech Acts by EFL Learners." TESOL Quarterly **27**(1): 33 - 56.

Cohen, A. D., Weaver, S. J. & Li, T. Y. (1995). The impact of strategies-based instruction on speaking a foreign language. Minneapolis, National Language Resource Centre, University of Minnesota.

Cohen, L., Manion, L. & Morrison, K. (2000). Research Methods in Education.

London, UK, Routledge/Falme.

Conacher, J. E. (2004). "Review of 'Conversation and Technology'." Language Learning & Technology **8**(1): 20 - 23.

Cook, T. D. & Campbell, D. T. (1979). Quasi-Experimantation: Design and Analysis Issues for Field Settings. Chicago, Rand McNally.

Cronbach, L. (1971). Test validation. Washington D.C., American Council on Education.

Cronbach, L. J. (1984). Essentials of psychological testing. New York, Harper & Row.

Cronbach, L. J. (1988). Five Perspectives on Validity Argument. Hillsdale, Lawrence Erlbaum.

Cronbach, L. J., Ed. (1989). A history of psychology in autobiography. Stanford, CA, Stanford University Press.

Cronbach, L. J. & Meehl, P. E. (1955). "Construct validity in psychological tests." Psychological Bulletin 52(281 - 302).

Crookes, G. (1990). "The Utterance, and Other Basic Units for Second Language Discourse Analysis " Applied Linguistics **11**(2): 183 - 199.

Cumming, A., Ed. (1995). Introduction: The Concept of Validation in Language Testing. Validation in Language Testing. Clevedon, Multilingual Matters Ltd.

Cumming, A., Grant, L., Mulcahy-Ernt, P. & Powers, D.E. (2004). "A teacher-verification study of speaking and writing prototype tasks for a new TOEFL." Language Testing **21**(2): 107 - 145.

Cutler, A. & Clifton, C. (1999). Comprehending spoken language: A blueprint of the listener.

Cziko, G. A. & Park, S. (2003). "Internet Audio Communication for Second language learning: A Comparative review of six programs." Language Learning & Technology 7(1): 15 - 27.

Dandonoli, P. & Henning, G. (1990). "An investigation of the construct validity of the ACTFL Proficiency Guidelines and Oral Interview procedure." Foreign Language Annals **23**(1): 11 -22.

Davidson, P. (2003). "Why technology has had only a minimal impact on testing in education." Retrieved October 2005, from http://www.squ.edu.om/ceto/etex2003/

Davies, A. (1977). The construction of language tests. Testing and Experiential Methods of Edinburgh Course in Applied Linguistics. J. D. Allen, A. Oxford, OUP. **4:** 38-104.

Davies, A. (1983). "The validity of concurrent validation." Development in Language Testing: 141-145.

Davies, A. (1984). "Validating three tests of English language proficiency." Language Testing(1): 50-69.

Davies, A. (1996). Outing the tester:Theoretical models and practical endeavours in language testing. Language and Education. G. M. Blue, R. Clevedon, Multilingual Matters. **11:** 60-69.

deBeer, M. & Visser, D. (1998). "Comparability of the paper-and-pencil and computerized adaptive versions of the General Scholastic Aptitude Test (GSAT) Senior." South African Journal of Psychology **28**(1): 21-27.

DeMaio, T. J. R., J.M. (1996). Cognitive Interviewing Techniques. In the lab and in the field. Answering questions: Methodology for cognitive and communicative processes in survey research. N. S. S. Sudman. San Francisco, CA, Jossey-Bass: 177-196.

Denzin, N. K. (1978). The research act. New York, McGraw Hill.

Dimitrova-Galaczi, E. (2004). Peer-peer Interaction in a Paired Speaking Test: The case of the First Certificate in English. Teachers College. Columbia, Collumbia University.

Dornyei, Z. (1995). "On the teachability of communication strategies." TESOL Quarterly 29: 55 - 85.

Dornyei, Z. & Kormos, J. (1998). "Problem-solving Mechanisms in L2 Communication: A Psycholinguistic Perspective." Studies in Second Language Acquisition **20**: 349 - 385.

Douglas, D. (1994). "Quantity and quality in speaking test performance." Language Testing **11**(2): 125-144.

Douglas, D. (2000). Assessing languages for specific purposes. Cambridge, CUP.

Douglas, D. & Selinker, L., Ed. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. A New Decade of Language Testing Research, TESOL.

Dreher, M. (1994). Qualitative research methods from the reviewer's perspective. Critical Issues in Qualitative Research Methods. J. Morse. London, UK, Sage.

Dudley-Evans, T. & St John, MJ. (1998). Developments in English for Specific Purposes. Cambridge, CUP.

Duff, P. A. (1986). Another look at interlanguage talk: Taking task to task Talking to Learn: Conversation in Second Language Acquisition R. R. Day. New York, Newbury House: 147 - 181.

Dunkel, P. A. (1991). "Computerized testing of non-participatory L2 listening comprehension proficiency: An ESL prototype development effort." Modern Language Journal 75(1): 64 - 73.

Dunkel, P. A. (1999). "Considerations in developing or using second/foreign language proficiency computer adaptive tests." Language Learning & Technology **2**(2): 77-93.

Egbert, J. (2004). "Review of 'Connected Speech'." Language Learning & Technology **8**(1): 24 - 28.

Eggins, S. & Slade, D. (1997). Analysing Casual Conversation. London, Cassell.

Egloff, B. & Schmukle, S. C. (2002). "Predictive validity of an Implicit Association test for assessing anxiety." Journal of Personality and Social Psychology **83**(6): 1441 - 1455.

Elder, C., Iwashita, N. & McNamara, T. (2001). "Can we predict task difficulty in an oral proficiency test? Exploring the potential of an Information-processing approach to task design." Language Learning **51**(3): 401-436.

Elder, C., Iwashita, N. & McNamara, T. (2002). "Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? ." Language Testing 19(4): 347-368.

Ellerton, A. (1997). Considerations in the validation of semi-direct oral testing. Reading, UK, University of Reading.

Ellis, R. J. (1980). "Oral Skills and their Identification." Reading English **14**(1): 31 - 34.

Ellis, N. C. (1996). "Sequencing in SLA: Phonological memory, chunking, and points of order." Studies in Second Language Acquisition **18**(1): 91-126.

Ellis, R. J. & Yuan, F. (2003). "The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production." Applied Linguistics **24**(1): 1 - 27.

Esling, J. H. (1991). Researching the Effects of Networking: Evaluating the Spoken and Written discourse generated by working with CALL. New York, Newbury House.

Evans, T. D. (1988). A consideration of the meaning of the word 'discuss' in examination questions. P. Robinson: 47 - 52.

Fetterman, D. M. (1998). "Empowerment evaluation and the internet: A synergistic relationship." Current Issues in Education **1**(4).

Field, J. (2003). Psycholinguistics: A resource book for students. London/New York, Routledge.

Fillmore, C. J. (1979). On Fluency, Academic Press Inc.

Foddy, W. (1993). Constructing questions for interviews and questionnaires. Cambridge, UK, CUP.

Fodor, J. A., Bever, T.G., and Garrett, M.F. (1974). The Psychology of Language. New York, McGraw Hill.

Foot, M. C. (1999). "Relaxing in pairs." ELT Journal **53**(1): 36 - 41.

Forster, K. (1974). The role of semantic hypothesis in sentence processing. Colloques Internationaux du CNRS No. 206, Paris.

Forster, K., Ed. (1976). Accessing the Mental Lexicon. New Approaches to Language Mechanisms. Amsterdam, Noth-Holland.

Forster, K., Ed. (1979). Levels of processing and the structure of the language processor. Sentence processing: Psycholinguistic studies presented to Merrill Garrett. New Jersey, LEA.

Foster, P., Tonkyn, A. & Wigglesworth, G. (2000). "Measuring spoken language: A unit for all reasons." Applied Linguistics **21**(3): 354-375.

Foster, P. & Skehan, P. (1996). "The influence of planning and task type on second language performance." Studies in Second Language Acquisition **18**: 299-323.

Foster, P. & Skehan, P. (1997). "Task type and task processing conditions as influences on foreign language performance." Langauge Teaching Research **1**(3): 185-211.

Foster, P. & Skehan, P. (1999). "The influence of source planning and focus of planning on task-based performance." Language Teaching Research **3**(3): 215-247.

Frase, L. T. (1996). "Technology for language assessment and learning." Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96: 519 - 543.

Fulcher, G. (1994). "Some priority areas for oral language testing." Language Testing Update **15**: 39-47.

Fulcher, G. (1996a). "Does thick description lead to smart tests?A data-based approach to rating scale construction." Language Testing **13**(2): 208-238.

Fulcher, G. (1996b). "Testing tasks: Issues in task design and the group oral." Language Testing **13**(1): 23-51.

Fulcher, G., Ed. (2000). Computers in language testing. A Special Interest in Computers. Manchester, IATEFL Publications.

Fulcher, G. (2003). "Interface design in computer-based language testing." Language Testing **20**(4): 384-408.

Fulcher, G. (2003). Testing second language speaking. London, Longman/ Pearson.

Fulcher, G. & Reiter, R. M. (2003). "Task difficulty in speaking tests." Langauge Testing **20**(3): 321-344.

Gaskell, M. G. & Marslen-Wilson, W. D. (1999). "Ambiguity, Competition, and Blending in Spoken Word Recognition." Cognitive Science **23**(4): 439 - 462.

Gasparro, J. E. (1986). Testing and teaching for oral proficiency: A familiarization kit. Boston, Heinle & Heinle.

Gass, S., & Varonis, E. M. (1985). Task variation and non-native/non-native negotiation of meaning. Input and Second language Acquisition. M. G. C. G. Madden. Rowley, MA, Newbury House: 149 - 161.

Gass, S. M. (1997). Modelling Second Language Acquisition(Ch 1); Input and second language acquisition theories (Ch 4); The role of interaction (Ch 5). Input, interaction and the second language learner. S. M. Gass. New Jersey, Lawrence Erlbaum Associates Publishers.

Gass, S. M. & Mackey, A. (2000). Stimulated recall methodology in second language research. New Jersey, Lawrence Erlbaum Associates.

Geisinger, K. F. (1992). "The metamorphosis of test validation." Educational Psychologists **27**: 197 - 222.

Geranpayeh, A. (2001). "CB BULATS: Examining the reliability of a computer-based test using test-retest method." UCLES EFL Research Notes **5**: 14-16.

Goodwin-Jones, B. (1997). "Real time audio and video playback on the web." Language Learning & Technology **1**(1): 5-8.

Goodwin-Jones, B. (1998). "New developments in digital video." Language Learning & Technology **2**(1): 11-13.

Goodwin-Jones, B. (2000). "Speech technologies for language learning." Language Learning & Technology **3**(2): 6-9.

Goodwin-Jones, B. (2001). "Language testing tools and technologies." Language Learning & Technology **5**(2): 8-12.

Goodwin-Jones, B. (2002). "Technology for prospective language teachers." Language Learning & Technology **6**(3): 10-14.

Gorard, S., & Taylor, C., Ed. (2004). Combining Methods in Education and Social Research. Conducting Educational Research. Maidenhead, Open University Press.

Grabe, W. & Kaplan, R. (1996). Theory and Practice of Writing. London, Longman.

Green, A. (1998). Verbal protocol analysis in language testing research. Cambridge, CUP.

Green, A. (2003). Test impact and English for Academic Purposes: A comparative study in backwash between IELTS preparation & University presessional courses. CRTEC. London, UK, University of Surrey Roehampton.

Guerrero, M. D. (2000). "The unified validity of the Four Skills Exam: applying Messick's framework." Language Testing 17 (4): 397-421.

Gumperz, J. J. (1972). Sociolinguistics and Communication in Small Groups. Harmondsworth, Penguin.
Gumperz, J. J. (1982). Discourse Strategies. Cambridge, Cambridge University Press.

Gutteridge, M. (2003). "Assistive technology for candidates with special needs." Research Notes 12: 15.

Gutteridge, M. (2006). "ESOL Special Circumstances 2004: a review of Upper Main Suite provision." Research Notes, Cambridge ESOL Examinations(23):17 - 19.

Halliday, M. A. K. (1975). Learning How to Mean, Edward Arnold.

Halliday, M. A. K. (1989). Two kinds of complexity. Oxford, Oxford University Press.

Hamilton, L. S., Nussbaum, E.M. & Snow, R.E. (1997). "Interviewing procedures for validating Science assessments." Applied Measurement in Education **10**(2): 181-200.

Hamp-Lyons, L. & Lynch, B. K., Ed. (1998). Perspectives on Validity: A Historical Analysis of Language Testing Conference Abstracts. Validation in Language Assessment. Mahwah, NJ, Lawrence Erlbaum Associates.

Hargreaves, R. & Fletcher, M. (1979). Making Polite Noises. London, Evans.

Hasselgren, A. (1996). "Oral test subskill scores: What they tell us about raters and pupils." Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96: 241 - 251.

Hasselgren, A. (2000). "A Messick-based system for Speaking test validation." Language Testing Update 27.

Hasselgren, A. (2002). Fluency and the small words of speaking. Assessing secondary school students' oral interaction. C. a. A. H. Estobar. Barcelona, Apac.

Hasselgren, A. (2004). Testing the Spoken English of Young Norwegians: a study of test validity and the role of «smallwords» in contributing to pupils' fluency. Cambridge, Cambridge University Press.

Hayes, J. R. & Hatch, J. A. (1999). "Issues in measuring reliability." Written Communication **16**(3): 354 - 367.

He, A. W. & Young, R. (1998). Language Proficiency Interviews: A Discourse Approach. Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency. R. Y. A. W. He. Amsterdam, Benjamins.

Henning, G. (1987). A guide to language testing. Cambridge, Mass, Newbury House.

Henning, G. (1991). Validating an Item Bank in a computer-assisted or computer-adaptive test: Using IRT for the process of validating CATS. New York, Newbury House.

Henning, G. (1996). "Accounting for non-systematic error in performance ratings." Language Testing **13**(1): 53 - 62.

Heubert, J. P. & Hauser, R. M. (1999). Tests as Measurements. High Stakes: Testing for Tracking, Promotion and Graduation, The National Academies Press: 71 - 85.

Hoey, M. (1992). Some properties of spoken discourse. Applied Linguistics and ELT. Review of ELT.

Honaker, L. M. (1988). "The equivalency of computerized and conventional MMPI administration: A critical review." Clinical Psychology Review **8**: 561-577.

Hubbard, C. (2003). "Feedback on CPE re-training." UCLES EFL Research Notes **12**: 16-17.

Huda, N., Ed. (1998). Relationship between speaking proficiency, reflectivity-impulsivity, and L2 learning strategies. Learners and Language Learning. Singapore, SEAMEO Regional Language Centre.

Hughes, A. (1989). Testing for Language Teachers. Cambridge, Cambridge University Press.

Hughes, A. (2003). Testing for Language Teachers. Cambridge, Cambridge University Press.

Hughes, R. (1996). English in Speech and Writing: Investigating Language and Literature. London, Routledge.

Hughes, R. (2002). Teaching and Researching Speaking. London, Pearson Education.

Hughes, R. (forthcoming). "Testing the visible: Literate biases in oral language testing."

Hymes, D., Ed. (1972). On Communicative Competence. Sociolinguistics. Harmondsworth, Penguin.

Hymes, D. (1985). "Toward linguistic competence." Revue de l'AILA: AILA Review **2**: 9-23.

Iwashita, N., Ed. (1999). Tasks and learners' output in NNS-NNS interaction (Japanese). Studies on the acquisition of Japanese as a second language. Amsterdam, John Benjamin.

James, G. (1983). "Innovation in oral examining-An eighteenth century spoken Latin test." Journal of The Royal Society of Arts.

James, G. (1989). Considerations in the design of an oral test in English for Academic Purposes. Exeter, UK, University of Exeter.

Jennings, M., Fox, J, Graves, B. & Shohamy, E. (1999). "The test-taker's choice: An investigation of the effect of topic on language-test performance." Language Testing **16**(4): 426-456.

Johnson, M. (2001). The art of non-conversation. Boston, Yale University Press.

Johnson, M. & Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? . Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency. R. Y. A. W. He. Amsterdam, Benjamins.

Jones, N. (2001). "Reliability as one aspect of test quality." UCLES EFL Research Notes **4**: 2-4.

Jones, N. (2003). "The role of technology in Langauge Testing." UCLES EFL Research Notes **12**: 3-4.

Jones, W. P. (1994). "Computer use and cognitive style." Jounal of Research on Computing in Education **26**(4): 514-521.

Jonson, J. L. & Plake, B. S. (1998). "A historical comparison of validity standards and validity practices." Educational and Psychological Measurement **58**(5): 736 - 753.

Kane, M. T. (1992). "An argument-based approach to validity." Psychological Bulletin 112: 527 - 535.

Kane, M. T. (2001). "Current concerns in validity theory." Journal of Educational Measurement **38**(4): 319-342.

Kane, M. T. (2002). "Validating high-stakes testing programs." Educational Measurement: Issues and Practice: 31-41.

Kane, M. T., Crooks, T. & Cohen, A. (1999). "Validating measures of performance." Educational Measurement: Issues and Practice: 5-17.

Kelly, R. (1978). On the construct validation of comprehension texts: An exercise in Applied linguistics, University of Queensland.

Kenyon, D., Ed. (1998). An Investigation of the Validity of Task Demands on Performance-based Tests of Oral Proficiency Validation in Language Assessment Long Beach, CA, Lawrence Erlbaum Associates.

Kenyon, D., Malabonga, V. & Carpenter, H. (2001). "Response to the Norris commentary." Language Learning & Technology **5**(2): 106-108.

Kenyon, D. & Malabonga, V., Ed. (1999). Multimedia computer technology and performance-based language testing: A demonstration of the Computerized Oral Proficiency Instrument (COPI). Computer Mediated language Assessment and Evaluation in Natural Language. New Brunswick, NJ: Association for Computational Linguistics.

Kenyon, D. & Malabonga, V. (2001). "Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments." Language Learning & Technology **5**(2): 60-83.

Kintsch, W. (1998). Comprehension: A Framework for Cognition. Cambridge, Cambridge University Press.

Kormos, J. (1999). "Simulating conversations in oral proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams." Language Testing **16**(2): 163-188.

Kramsch, C. (1986). "From language proficiency to interactional competence." The Modern Language Journal 70(4): 366 - 372.

Krueger, R. A. (2002). "Designing and Conducting Focus Group Interviews." 2003.

Kunan, A. J. (1995). Test taker characteristics and test performance: A structural modeling approach. Cambridge, CUP.

Kuo, J., & Jiang, X. (1997). "Assessing the assessments: The OPI and the SOPI." Foreign Language Annals **30**(4): 503-512.

Lantolf, J. P. & Frawley, W. (1985). "Oral proficiency testing: A critical analysis." The Modern Language Journal **69**(4): 337-345.

Lantolf, J. P. & Frawley, W. (1988). "Proficiency: Understanding the Construct " Studies in Second Language Acquisition **10**: 181 - 195.

Lazaraton, A. L. (1992). A conversation analysis of structure and interaction in the language. Ann Arbor, MI, USA, University of Ann Arbor, MI.

Lazaraton, A. L. (1992). "The structural organization of a language interview:A conversational analytic perspective." System **20**(3): 373-386.

Lazaraton, A. L. (1996a). "Interlocutor support in oral proficiency interviews: the case of CASE." Language Testing **13**(2): 151-172.

Lazaraton, A. (1996b). A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE). 15th Language Testing Research Colloquium. Cambridge & Arnhem, Cambridge University Press.

Lazaraton, A. L. (2000). An analysis of the relationship between task features and candidate output for the revised IELTS Speaking test. Cambridge, UK, Division of EFL UCLES: 1-30.

Lazaraton, A. L. (2001). Qualitative research methods in language test development and validation. ALTE, Barcelona.

Lemke, J. L. (1998). Analysing verbal data: Principles, methods and problems. International Handbook of Science Education, Kluwer Academic Publishers: 1175-1189.

Lennon, P. (1990). "Investigating fluency in EFL: A quantitative approach." Language Learning **40**(3): 387 - 417.

Lessler, J. T. F., B.H. A coding system for appraising questionnaires. Answering Questions: 259-291.

Levelt, W. J. M. (1989/1993). Speaking: From intention to articulation. Cambridge, MA, MIT Press.

Levelt, W. J. M., Praamstra, P., Meyer, A.S., Helenius, P., & Salmelin, R. (1998). "An MEG study of picture naming." Journal of Cognitive Neuroscience 10: 553-567.

Levelt, W. J. M. (1999). Producing spoken language:A blueprint of the speaker. The Neurocognition of Language. C. H. Brown, P.

Linacre, J. M. (1989). Many-Facet Rasch Measurement. Chicago, MESA Press.

Long, M. H. (1981). "Questions in foreigner talk discourse." Language Learning 31(1).

Long, M. H. (1989). "Task, group, and task-group interactions." University of Hawai'i Working Papers in ESL **8**(2): 1 - 26.

Long, M. H. & Porter, P. A. (1985). "Group work, Interlanguage talk, and Second language acquisition." TESOL Quarterly **19**(2): 207 - 228.

Lowe, P. J., Ed. (1981). Structure of the oral interview and content validity. The Construct Validation of tests of Communicative Competence. Washington DC, TESOL.

Lowe, P. J. (1983). "The ILR oral interview: Origins, applications, pitfalls, and implications." Die Unterrichtspraxis **60**: 230 - 44.

Lowe, P. J. & Clifford, R. T., Ed. (1980). Developing an indirect measure of overall oral proficiency.

Lu, Y. (2003). "Insights into the FCE Speaking test." UCLES EFL Research Notes **11**(Feb 2003): 15-19.

Luecht, R. M. (2001). New directions in computerized testing research. Language Testing Research Colloquium. St Louis, MO.

Lumley, T. & O'Sullivan, B. (2005). "The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking." Language Testing 22(4): 415 - 437.

Lumley, T. & Brown, A. (1997). "Interlocutor variability in specific-purpose language performance tests." (in A. Huhta et al.): 137-50.

Luoma, S. (2004). Assessing Speaking. Cambridge, Cambridge University Press.

Lynch, B. K. & McNamara, T.F. (1998). "Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants " Language Testing 15(2): 158-180.

Lynch, T. & K. Anderson (2001). "The Value of an Additional Native Speaker in the English Language Classroom." Edinburgh Working Papers in Applied Linguistics(11): 69-80.

Mackey, A., Gass, S. & McDonough, K. (2000). "How do Learners Perceive Interactional Feedback." Studies in Second Language Acquisition **22**: 471 - 497.

Madsen, H. S. (1983). Techniques in Testing. Oxford, Oxford University Press.

Madsen, H. S., & Jones, R. L. (1981). Classification of oral proficiency tests. Papers in Language Testing 1967-74 L. P. B. Spolsky. Washington DC, TESOL: 15-30.

Malabonga, V., Kenyon, D. M. and Carpenter, H. (2005). "Self-assessment, preparation and response time on a computerized oral proficiency test." Language Testing **22**(1): 59 - 92.

Malaysia, E. (2001). "Developing Malaysia into a Knowledge-based Economy." Retrieved January 2006, 2006, from http://www.epu.jpm.my.

Malone, M. E. (1999). The development of the English speaking test: An investigation of reliability and validity. Washington, D.C., Georgetown University.

Marslen-Wilson, W. D., & Tyler, L.K. (1980). "The temporal structure of spoken language understanding." Cognition **8**: 1 -71.

Marslen-Wilson, W. D., Tyler, L.K., & Seidenberg, M., Ed. (1978). Sentence Processing and the Clause Boundary. Studies in Sentence Perception. New York, Wiley.

McCarthy, M. (1998). Spoken Language & Applied Linguistics. Cambridge, Cambridge University Press.

McCarthy, M. & Carter, R. (1994). Language as Discourse: Perspectives for Language Teaching. London, Longman.

McDonald, A. S. (2002). "The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments." Computers and Education **39**: 299-312.

McNamara, T. (1996). Measuring second language performance. Harlow, Longman.

McNamara, T. (1997). "'Interaction' in second language performance assessment: Whose performance?" Applied Linguistics **18**(4): 446-466.

McNamara, T. (2000). Language Testing. Oxford, Oxford University Press.

McNamara, T. (2003). Validity and reliability in the senior school curriculum: New takes on old questions. Australasian Curriculum, Assessment and Certification Authorities (ACACA) National Conference, Adelaide.

McNamara, T., & Adams, J. (1996). "New approaches to the analysis of task- and rater-related dependencies in performance assessments." Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96: 625 - 635.

Mead, A. D. & Drasgow., F. (1993). "Equivalence of computerized paper-and-pencil cognitive ability tests: A meta-analysis." Psychological Bulletin **114**(3): 449-458.

Meier, S. T. (2000). "Consistency across constructs: Lack of discriminant validity." Chapter 5. Retrieved 22 April, 2004, from http://www.acsu.buffalo.edu/~stmeier/c5.html.

Meiron, B. E. & Schick, L.S. (2000). Ratings, raters and test performance: An exploratory study. Fairness and validation in language assessment. A. J. Kunan. Cambridge, UK, CUP.

Messick, S. (1975). "The standard program: Meaning and values in measurement and evaluation." American Psychologist 30: 955 - 966.

Messick, S. (1980). "Test Validity and the Ethics of Assessment." American

Psychologist **35**(11): 1012-1027.

Messick, S. (1982). "Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing policy." Educational Psychologists **17**(2): 67 - 91.

Messick, S. (1988). The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement. Test Validity. H. B. Wainer, H. Hillsdale, New Jersey, Erlbaum: 35-45.

Messick, S. (1989). "Meaning and values in test validation: The science and ethics of assesment." Educational Researcher **18**(2): 5-11.

Messick, S. (1989). Validity. Educational Measurement. R. Linn. New York, Macmillan: 13-103.

Messick, S. (1992). Validity of test interpretation and use. Encyclopedia of Educational Research. M. C. Alkin. New York, Macmillan.

Messick, S. (1995). "Standards of validity and validity of standards in performance assessment." Educational Measurement: Issues and Practice **14**(4): 5-8.

Messick, S. (1996). "Validity and washback in language testing." Language Testing **13**(3): 241-256.

Milanovic, M. & Saville, N. (1996). Performance testing, Cognition and Assessment. Cambridge, CUP.

Miller, S. I. & Marcel, F. (1994). Qualitative Research Methods: Social Epistomology and Practical Inquiry. New York, Peter Lang.

Mislevy, R. J., Steinberg, L. S. and Almond, R. G. (2003). "On the structure of assessment arguments." Measurement: Interdisciplinary Research and Perspectives 1(1): 3 - 62.

Moller, A. D. (1982). A study in the validation of proficiency tests of English as a foreign language, University of Edinburgh.

Moore, T. & Morton, J. (1999). Authenticity in the IELTS Academic Module Writing Test:A comparative study of Task 2 items and university assignments. Tulloh: 64 -106.

Moss, P. A. (1992). "Shifting conceptions of validity in educational measurement: Implications for performance assessment." Review of Educational Research(62): 229-258.

Moss, P. A. (1994). "Can there be validity without reliability?" Educational Researcher **23**: 5-12.

Mullen, K. A., Ed. (1978). Determining the effect of uncontrolled sources of error in a direct test of oral proficiency and the capability of the procedure to detect improvement following classroom instruction. Direct Tests of Speaking Proficiency: Theory and Application. Princeton, Educational Testing Service.

Murphy, E. (2004) Promoting Construct Validity in Instruments for the Analysis of Transcripts of Online Asynchronous Discussions. Educational Media International **Volume**, 347 - 354 DOI:

Neuman, G. & Baydoun, R. (1998). "Computerization of paper-and-pencil tests." Applied Psychological Measurement **22**(1): 71-83.

Norris, J., Brown, J. D., Hudson, T. & Yoshioka, J. (1998). Designing second language performance assessments. Hawai'i, University of Hawai'i Press.

Norris, J. M. (1997). "The German Speaking Test: Utility and caveats." Die Unterrichtspraxis/Teaching German 30(2): 148-158.

Norris, J. M. (2001). "Concerns with computerized adaptive oral proficiency assessment." Language Learning & Technology **5**(2): 99-105.

Norris, J. M. & Ortega, L. (2000). "Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis." Language Learning 50: 417-528.

North, B. & Schneider, G. (1998). "Scaling descriptors for language proficiency scales." Language Testing **15**(2): 217-263.

Nteliou, E. (2000). UCLES 'Main Suite' Speaking tests: Describing the test-takers' outputin terms of CALS checklist of operations at KET and FCE levels. Dept of Linguistic Science. Reading, UK, University of Reading.

O'Loughlin, K. (1997). The Comparability of a Direct and Semi-direct Speaking Tests: A Case Study. Melbourne, University of Melbourne.

O'Loughlin, K., Ed. (1997). Test-taker performance on direct and semi-direct versions of the oral interaction module. ACCESS: Issues in English language test design and delivery. Melbourne.

O'Loughlin, K. (2001). The equivalence of direct and semi-direct speaking tests. Cambridge, UK, CUP.

O'Loughlin, K. (2002). "The impact of gender in oral proficiency testing." Language Testing **19**: 169-192.

O'Sullivan, B. (1995). Oral language testing: Does the age of the interlocutor make a difference?, University of Reading.

O'Sullivan, B. (2000b). "Exploring gender and oral proficiency interview performance." System **28**: 373-386.

O'Sullivan, B. (2000a). Towards a model of performance in oral language testing. Centre for Applied Language Studies. Reading, UK, University of Reading.

O'Sullivan, B. (2001). Developing tests of Speaking. Budapest, ALTE.

O'Sullivan, B. (2001). Elements of uncertainty: Defining characteristics of the test taker. Nottingham, Language Testing Forum.

O'Sullivan, B. (2002). "Learner acquaintanceship and oral proficiency test pair-task performance." Language Testing **19**(3): 277-295.

O'Sullivan, B., Ed. (2004). Modelling factors affecting oral language test performance: A large scale empirical study. Studies in Language Testing. Cambridge, CUP.

O'Sullivan, B. (2005). The Critical Role of Assessment in Learning. Current Trends in English Language Testing (CTELT), Dubai.

O'Sullivan, B., Ed. (2006). Issues in Testing Business English. Studies in Language Testing. Cambridge, Cambridge University Press.

O'Sullivan, B., & Porter, D. (1995). The importance of audience age for learner-speakers and learner-writers from different cultural backgrounds. RELC. Singapore.

O'Sullivan, B. & Porter, D. (1998). The effect of learner acquaitanceship on pair-task performance. Language Testing Research Colloquium, Monterey, CA.

O'Sullivan, B., Porter, D. & Weir, C.J. (1999). Research issues in testing spoken language. Cambridge, UCLES.

O'Sullivan, B., Saville, N. & Weir, C.J. (2002). "Using observation checklists to validate speaking test tasks." Language Testing **19**(1): 33-56.

O'Sullivan, B. a. Weir, C. (2002). Research issues in testing spoken language. Cambridge, Cambridge ESOL.

O'Sullivan, B., Weir, C. J. and Horai, T. (2004). Exploring difficulty in speaking tasks: An intra-task perspective. Cambridge, ESOL University of Cambridge.

Olson, D. R. (1977). "From Utterance to Text: The Bias of Language in Speech and Writing." Harvard Educational Review **47**(3): 257 - 281.

Oomen, C. E. & Postma, A. (2002). "Limitations in Processing Resources and Speech Monitoring." Language & Cognitive Processes **17**(2): 163-84.

Oppenheim, A. N. (2000). Questionnaire design, Interviewing and Attitude measurement Continuum International Publishing Group - Academi

Ortega, L. (1999). "Planning and focus on form in L2 oral performance." Studies in Second Language Acquisition **21**: 109-148.

Patri, M. (2002). "The influence of peer feedback on self and peer-assessment of oral skills." Language Testing **19**(2): 109-131.

Paul, J. W. (1994). "Computer use and cognitive style." Retrieved March 2003.

Pavlou, P. (1996). "Do different speech interactions in an oral proficiency test yield different kinds of language?" Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96: 185 - 201.

Pendergast, T. M. (1985). OLAF N.73: A computerized oral language analyzer and feedback system. New Directions in Language Testing. Y. P. Lee, Fok, A., Lord, R. & Low, G., Pergoman Press Ltd.

Perrett, G. (1987). The language testing interview: A reappraisal 8th World Congress of Applied Linguistics. Sydney.

Pica, T., Lincoln-Porter, F., Paninos, D. & Linnel, J. (1996). "Language learners'

interaction: How does it address the input, output, and feedback needs of L2 learners?" TESOL Quarterly **30**(1): 59 - 84.

Porter, D. (1991a). Affective factors in language testing. Language Testing in the 1990s. J. C. N. Alderson, B. London, Modern English Publication: 32-40.

Porter, D., Ed. (1991b). Affective factors in the assessment of oral interaction: Gender and status. Current Developments in Language Testing. Singapore, SEAMEO Regional Language Centre.

Porter, D. & Shen, S. H., Ed. (1991). Sex, status and style in the interview. The Dolphin 21. Aarhus, Aarhus University Press.

Porter, P. A. (1986). How learners talk to each other: Input and interaction in task-centered discussions. Talking to Learn: Conversation in Second Language Acquisition. R. R. Day. New York, Newbury House: 200 - 222.

Poulisse, N. (1990). The Use of Compensatory Strategies by Dutch Learners of English. Berlin, Mouton de Gruijter.

Powers, D. E., Schedl, M. A., Leung, S. W., & Butler, F. A. (1999). "Validating the revised Test of Spoken english against a criterion of communicative success." Language Testing **16**(4): 399 - 425.

Pridham, F. (2001). Language of Conversation (Linguistic Theory Guides). New York, Routledge.

Purpura, J. E., Ed. (1999). Learner strategy use and performance on language tests: A structural equation modelling approach. Studies in Language Testing. Cambridge, CUP.

Qualls, A. L. a. M., A. D. (1996). "The degree of congruence between test standards and test documentation within journal publications." Educational and Psychological Measurement **56**: 209 - 214.

Raffaldini, T. (1988). "The use of situation tests as measures of communicative ability." Studies in Second Language Acquisition **10**: 197 - 216.

Rea-Dickins, P. G., K. (1992). Evaluation. Oxford, OUP.

Reeves, T., Ed. (1991). From testing research to educational policy: A comprehensive test of oral proficiency. Language Testing in the 1990s: the

Communicative Legacy. London, Macmillan.

Rethinasamy, S. (2005). The Effects of Different Rater Training Procedures on ESL Essay Rating Judgement, Roehampton University London.

Riggenbach, H. (1998). Evaluating Learner Interactional Skills: Conversation at the Micro Level. Talking and Testing: Discourse Approaches to the Assessment of oral Proficiency. R. Y. A. W. He.

Robinson, P. (1995). "Task complexity and second language narrative discourse." Language Learning **45**(1): 99-140.

Robson, C. (2002). Real World Research. Oxford, Blackwell Publishing.

Roever, C. (2001). "Web-based language testing." Language Learning & Technology **5**(2): 84-94.

Roever, C. (2001). A web-based test of interlanguage pragmalinguistic knowledge: Speech acts, routines, implicatures. Graduate Division, University of Hawai'i, USA.

Ross, S. (1992). "Accomodative questions in oral proficiency interviews." Language Testing **9**: 173 - 186.

Ross, S. & Berwick, R. (1992). "The discourse of accomodation in oral proficiency." Studies in Second Language Acquisition **14**: 159-176.

Rulon, K. A. & McCreary, J. (1986). Negotiation of content: Teacher-fronted and small-group interaction. Talking to Learn: Conversation in Second Language Acquisition R. R. Day. New York, Newbury House: 182 - 199.

Russell, M. (1999). "Testing on computers: A follow-up study comparing performance on computer and on paper." Education Policy Analysis Archives 7(20): 1-48.

Russell, M., & Haney, W. (2000). "Bridging the gap between testing and technology in schools." Education Policy Analysis Archives **8**(19): 1-10.

Russell, M. & Haney, W. (1997). "Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil." Education Policy Analysis Archives **5**(3).

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. Educational Measurement: Issues and Practice: 7-15.

Sacks, H., Schegloff, E. and Jefferson, G. (1974). "A simplest systematics for the organization of turn-taking for conversation " Language **50**: 696 - 735.

Salvia, J. Y., James E. (2003). "A brief explanation of Item-response theory." Retrieved 16/10/2003, from http://college.hmco.com/education/assessment.

Schiffrin, D. (1994). Approaches to discourse. Oxford, Blackwell Publishers.

Schrank, F. A., Fletcher, T. V., and Alvarado, C. G. (1996). "Comparative validity of three English oral proficiency tests." The Bilingual Research Journal **20**(1): 55 - 68.

Shannon, D. M., Johnson,T. E., Searcy, S., and Lott, A. (2002) Using electronic surveys: advice from survey professionals. Practical Assessment, Research & Evaluation **Volume**, DOI:

Shaw, S. (2002). "The effect of training and standardization on rater judgement and inter-rater reliability." UCLES EFL Research Notes **8**: 13-17.

Shepard, L. A., Ed. (1993). Evaluating test validity. Review of Research in Education. Washington, DC, American Educational Research Association.

Shepard, L. A. (1997). "The Centrality of test Use and Consequences for Test Validity." Educational Measurement: Issues and Practice **16**(2): 5-8, 13, 24.

Shermis, M. D., & Lombard, D. (1998). "Effects of computer-based test administrations on test anxiety and performance." Computers in Human Behaviour **14**(1): 111-123.

Shohamy, E. (1983). "The stability of oral language proficiency assessment on the oral interview testing procedure." Language Learning **33**: 527-540.

Shohamy, E. (1988). "A proposed framework for testing the oral language of second/foreign language learners." Studies in Second Language Acquisition **10**: 165-179.

Shohamy, E. (1994). "The validity of direct versus semi-direct oral tests." Language Testing **11**: 99-123.

Shohamy, E. (1996). "Test impact revisited: Washback effect over time." Language Testing **13**(3): 298-317.

Shohamy, E., Reeves, T. and Bejarano, Y. (1986). "Introducing a new comprehensive test of oral proficiency." English Language Teaching Journal **40**: 212 - 222.

Shohamy, E. & Stansfield, C. W. (1990). "The Hebrew Speaking test: An example of international cooperation in test development and validation " AILA Review 7: 79 - 90.

Sigman, J. S. & Donnellon, A. (1989). Discourse rehearsal: Interaction simulating interaction. Communication and Simulation. D. S. Crookall, D., Multilingual matters Ltd.: 69-81.

Skehan, P. (1996). "A framework for the implementation of task-based instruction." Applied Linguistics **17**: 38-62.

Skehan, P. (1998). A cognitive approach to language learning. Oxford, OUP.

Skehan, P. (1998b). "Processing perspectives to second language development, instruction, performance and assessment " Thames Valley Working papers in Applied Linguistics 4: 70 - 88.

Spolsky, B. (1968). "Language Testing - The problem of validation." TESOL Quarterly 2: 88-94.

Spolsky, B., Ed. (1978). Introduction: Linguists and language testers. Advances in Language testing Research: Approaches to Language testing. Washington DC, Centre for Applied Linguistics.

Spolsky, B. (1989). "Communicative competence, language proficiency, and beyond." Applied Linguistics **10**(2): 138-156.

Spolsky, B. (1990). "Oral examinations: an historical note." Language Testing 7(2): 158-173.

Spolsky, B. (1995). Measured Words. Oxford, OUP.

Staff, E. A. (1997). "Designing structured interviews for educational research."

Stansfield, C. W. (1989). Simulated Oral Proficiency Interviews. Washington DC. Centre for Applied Linguistics.

Stansfield, C. W. (1991). A comparative analysis of simulated and direct oral proficiency interviews. Singapore, Regional English Language Center: 199-209.

Stansfield, C. W. & Kenyon, D.M. (1988). Development of the Portuguese Speaking Test. Washington, DC, Centre for Applied Linguistics.

Stansfield, C. W. & Kenyon, D.M. (1989). Development of the Hausa, Hebrew and Indonesian Speaking Tests. Washington, DC, Centre for Applied Linguistics.

Stansfield, C. W. & Kenyon, D. M. (1991). Development of the Texas Oral Proficiency Test (TOPT). Washington, DC, Center for Applied Linguistics.

Stansfield, C. W. & Kenyon, D.M. (1992). "Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview." System **20**(347-364).

Stansfield, C. W. & Kenyon, D.M. (1996). "Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines." Modern Languages in Practice **2**: 124-153.

Stansfield, C. W. & Wu, W. M. (2001). "Towards authenticity of task in test development." Language Testing 18(2): 187 - 206.

Strauss, A. & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. Newbury Park, Sage.

Stricker, L. J. (2004). "The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test." Language Testing **21**(2): 146 -173.

Suhua, H. (1998). A communication test of spoken English for the CET6. Shanghai, ROC, Shanghai Jiao Tong University.

Swain, M. (2001). "Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores." Langauge Testing **18**(3): 275 - 302.

Swain, M., and Lapkin, S. (1998). "Interaction and second language learning: Two adolescent French immersion students working together." Modern Language Journal **82**: 320 - 337.

Tannen, D. (1989). Talking voices: Repetition, dialogue, and imagery in conversational discourse. Cambridge, Cambridge University Press.

Tannen, D. (1996). Talking from 9 to 5: Women and Men at Work - Language, Sex and Power London, Virago Press Ltd

Tarone, E. (1998). Research in interlanguage variation: Implications for language testing. Interfaces between second language acquisition and language testing research. L. F. C. Bachman, A.D. Cambridge, Cambridge University Press.

Taylor, L. (2001). "Revising the IELTS Speaking test: Developments in test format and task design." UCLES EFL Research Notes **5**: 2-5.

Taylor, L. (2001). "Revising the IELTS Speaking test: Retraining IELTS examiners worldwide." UCLES EFL Research Notes **6**: 9-11.

Taylor, L. & Gutteridge, M. (2003). "Responding to diversity: providing tests for langugae learners with disabilities." Research Notes 11: p.15.

Taylor, L. & Jones, N. (2001). "Revising the IELTS Seaking test." UCLES EFL Research Notes **4**: 9-12.

Taylor, L. & Shaw, S. (2002). "CELS Speaking:test development and validation activity." UCLES EFL Research Notes **9**: 13-15.

Thelwall, M. (2000). "Computer-based assessment:: A versatile educational tool." Computers and Education **34**: 37-49.

Thomson, P., Saunders J. & Foyster, J. (2001). Improving the validity of competency-based assessment. Australia, National Centre for Vocational Education Research.

Tonkyn, A. (1996). The oral language development of instructed second language learners: The quest for a progress sensitive proficiency measure. Change and Language. H. C. a. L. Cameron. Clevedon: 116 - 130.

Tyler, L. K., and Marslen-Wilson, W.D. (1977). "The on-line effects of semantic context on syntactic processing." J. verb. Learn. verb. Behave. **16**: 683 - 692.

UCLES (2002). "Review of recent validation studies." UCLES EFL Research Notes **10**: 22-23.

Underhill, N. (1987). Testing spoken language: A handbook of oral testing techniques. Cambridge, Cambridge University Press.

Upshur, J. A. & Turner, C. (1995). "Constructing rating scales for second language tests." ELT Journal **49**(1): 3-12.

Upshur, J. A. & Turner, C. (1999). "Systematic effects in the rating of second language speaking ability: Test method and learner discourse." Language Testing **16**(1): 82-111.

Ur, P. (1996). A Course in Language Teaching: practice and theory. Cambridge, Cambridge University Press.

Urquhart, A. H. & Weir, C. J. (1998). Reading in a Second Language: Process, Product and Practice. Harlow, Longman.

Valenti, S., Cucchiarelli, A. & Panti, M. (2001). "A framework for the evaluation of test management systems." Current Issues in Education **4**(6).

Van Lier, L. (1989). "Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversation." TESOL Quarterly **23**: 489-508.

Van Lier, L. (1995). Introducing language awareness. Harmondsworth, Middlesex, Penguin.

VanPatten, B. (1990). "Attending to content and form in the input: An experiment in consciousness." Studies in Second Language Acquisition 12: 287 - 301.

Wall, D. & Alderson, J.C. (1993). "Examining washback: The Sri Lankan impact study." Language Testing **10**: 41-69.

Weigle, S. C. (1998). "Using FACETS to model rater training effects." Language Testing **15**(2): 263-287.

Weigle, S. C. (2002). Assessing Writing. Cambridge, Cambridge University Press.

Weigle, S. C. & Lynch, B. (1996). "Hypothesis testing in construct validation." Modern Languages in Practice **2**: 58-71.

Weir, C. J. (1983a). Identifying the Language Problems of Overseas Students in Tertiary Education in the UK, University of London.

Weir, C. J. (1988). The specification, realization and validation of an English language proficiency test. Testing English for University Study. A. Hughes. Hong Kong, Modern English Publications: 45 -110

Weir, C. J. (1990). Communicative Language Testing New York, Prentice Hall.

Weir, C. J. (1993). Understanding and developing language tests, Prentice Hall.

Weir, C. J. (1994). Evaluation in ELT. Oxford, Blackwell.

Weir, C. J. & Milanovic, M. (2003). A survey of the history of the Certificate of Proficiency in English (CPE) in the twentieth century. Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913 - 2002. Cambridge, Cambridge University Press.

Weir, C. J., Ed. (2005). Language Testing and Validation: An Evidence-Based Approach. Research and Practice in Applied Linguistics. Basingstoke, Palgrave Macmillan.

Weir, C. J. (2005). "Limitations of the Common European Framework for developing comparable examinationsand tests." Language Testing **22**(3): 1 - 20.

Weir, C. J. (2005b). A Socio-Cognitive Approach to Test Validation. 8th MELTA Biennial International Conference, Malaysia.

Widdowson, H. G. (1983). Learning purpose and language use. Oxford, Oxford University Press.

Widdowson, H. G. (1989). "Knowledge of Language and Ability for Use." Applied Linguistics **10**(2): 128 -137.

Wigglesworth, G. (1997). "An investigation of planning time and proficiency level on oral test discourse." Language Testing **14**(1): 85-105.

Wigglesworth, G. & O'Loughlin, K. (1993). "An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English." Melbourne Papers in Language Testing **2**(1): 56 - 67.

Wilds, C. P. (1975). The oral interview test. Testing language proficiency R. L. J. B. Spolsky. Washington, DC, Center for Applied Linguistics: 29-38.

Wilkinson, A. S., L. (1969). "The evaluation of spoken language." Educational Review: 183-195.

Willis, J. (2000). Lexical chunks: Identifying frequent phrases, classifying and teaching them. Association for Language Awareness 2000. University of Leicester.

Wolfson, N. (1986). "Research methodology and the question of validity." TESOL Quarterly **20**(4): 689 - 699.

Wylie, E. & Ingram, D. E., Ed. (1993). Assessing speaking proficiency in the International English Language Testing system. A New Decade of Language Testing Research, TESOL.

Young, R. (2002). "Discourse approaches to oral language assessment." Annual Review of Applied Linguistics 22: 243 - 262.

Young, R. & Milanovic, M. (1992). "Discourse variation in oral proficiency interviews." Studies in Second Language Acquisition **14**: 440-442.

Zandvliet, D. & Farragher, P. (1997). "A comparison of computer-administered and written tests." Jounal of Research on Computing in Education **29**(4): 423-438.

Zimmerman, D. H. & West, C. (1973). Sex roles, interruptions and silences in conversation. Language and sex: Difference and dominance B. T. N. Henley. Rowley, MA, Newbury House.

APPENDICES

	Page
Appendix 3.1: Questionnaire difficulty & comments	A1-A7
Appendix 3.2: Student questionnaire (direct test)	A8-A14
Appendix 3.3: Staff questionnaire (direct test)	A15-A23
Appendix 3.4: Student Interview Notes	A24-A34
Appendix 3.5: Staff Interview Notes	A35-A45
Appendix 3.6: Guidelines for the Interview	A46-A50
Appendix 3.7: Speaking test documents	A51
a) Course syllabus	
b) Test specifications	
c) Sample question paper	
d) Score sheets	
e) Instructions for the speaking test and assessment procedure	
f) Criteria/Rating scale	
Appendix 3.8: Computer test script Task B	A52-A53
Appendix 3.9: Computer test questionnaire: CV and TBV	A54-A60
Appendix 3.10: Report workshop May2005	A61-A69
Appendix 3.11: Computer speaking test script A & B	A70-A71
Appendix 3.12: Computer test specifications	A72-A79

APPENDICES IN CD I

Appendix 3A: Validity evidence table
Appendix 3B: Pilot data March 2003 & pilot data September 2003
Appendix 3C: Extract from March 2003 pilot data SPSS output
Appendix 3D: The monologue test
Appendix 3E: Developing criteria for rating a speaking test
Appendix 3F: Staff questionnaire SPSS data (May 2005 workshop)
Appendix 3G: Rating the direct test (using old scale and TOEFL)
Appendix 3H: Main Study 2 findings (Pilot + MS2)
Appendix 3I: Student computer familiarity data
Appendix 3J: Rating the direct test + computer test (September 2005)
Appendix 3K: IRR direct test and computer test (September 2005)
Appendix 3L: Student questionnaire data on computer test & direct test (SPSS)
Appendix 3M: Main Study 1 findings summary
Appendix 3N: Transcripts student and staff interview Main study 1

APPENDICES IN CD II: The computerized speaking test

1. in Power Point
2. in WAVE Audio files (Task A + Task B)

APPENDIX 3.1

Attached is the questionnaire which I have marked (in red) the difficult words/phrases. Also some comments/feedback from both groups (high and low levels
Hope I did the right thing and this helps.

-ROS-

High proficiency group:

1. Too many items in Section A
2. The items are easy to understand but some are quite confusing
3. Some words have similar meaning

Low proficiency group:

1. The questionnaire is difficult/confusing
2. Too many items
3. Some difficult words/ unfamiliar
4. Do not understand a few sentences (meaning of the whole sentence)
   E.g. Section A: Items no.41 and 42 (a)
   Section B. Items no 10 and 26

5. Sentences quite long

APPENDIX 3.1

SPEAKING TEST: STUDENT QUESTIONNAIRE

Dear student:

We are conducting research on the university speaking test. This questionnaire aims to gather data as part of this research project. We want to get your feedback on the existing speaking test.

The English Department of Pusat Bahasa conducts the speaking test every semester for students who are in the 2nd and 3rd semesters of their diploma programmes; the codes are BEL 200 and BEL 250 respectively. It is a face-to-face test where candidates work with at least two examiners. This test was designed based on the MUET (Malaysian University English Test, developed by the Malaysian Examinations Council) syllabus and it includes two major tasks of

a) an individual presentation
b) a group discussion.

The items in this questionnaire are divided according to various aspects of the test. You will thus respond to questions pertaining to the content of the test, what you did to perform the tasks and how you think the test affects speaking and learning.

The data collected from you will help us improve the existing test. Therefore, your cooperation in the matter is greatly appreciated.

Thank you.

Sincerely,

Saidatul A Zainal abidin
Pusat Bahasa
UiTM Shah Alam

Name: **Student number:**
Gender:
Faculty:

This questionnaire has 3 sections. One is on the content of the test, one on what you have done in performing each task and one on the effects of the test on teaching and learning.

For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on a five point scale where:

1= Strongly disagree, 2= Disagree, 3= Undecided, 4= Agree, 5= Strongly agree

SECTION A: TEST CONTENT

1. Task A clearly states what I am required to do in response to situation.	1 2 3 4 5
2. Task B clearly states what I am required to do in the group discussion.	1 2 3 4 5
3. The individual presentation format is an appropriate test of my academic speaking skills.	1 2 3 4 5
4. The individual presentation format is an appropriate test of my ability. to speak English in social situations.	1 2 3 4 5
5. The group discussion format is appropriate for testing my ability to communicate orally in an academic context.	1 2 3 4 5
6. The group discussion format is appropriate for testing my ability to communicate in a social interaction.	1 2 3 4 5
7. Both the presentation and the group discussion should have equal marks.	1 2 3 4 5
8. I fully understand the three criteria for marking tasks A and B, i.e. **task fulfillment, language, and communicative ability.**	1 2 3 4 5
9. All three criteria are equally important.	1 2 3 4 5
10. There are other criteria that should be used in marking the test. If you agree, list your suggestions below.	1 2 3 4 5

_____

_____

11. The criteria for rating are clear to all candidates. 1 2 3 4 5

12. The order of the tasks, i.e. task A followed by task B consecutively is appropriate for the test. 1 2 3 4 5

13. Two minutes is sufficient time for a candidate to demonstrate his/her ability to present ideas through reasoning and explaining. 1 2 3 4 5

14. Ten minutes is sufficient time for a group of four candidates to demonstrate their ability to conduct a discussion on why a decision had been made. 1 2 3 4 5

15. Having written instructions to prepare for tasks A and B is helpful. 1 2 3 4 5

16. The instructions give sufficient information for candidates to prepare for tasks A and B. 1 2 3 4 5

17. The following would be more helpful ways of preparing for Tasks A and B:

a) Listen to a recorded presentation (lecture, dialogue, announcement, etc.) 1 2 3 4 5

b) Watch a video presentation 1 2 3 4 5

c) Look at visual text (pictures, graphs, maps, brochures, etc.) 1 2 3 4 5

d) The examiner gives you the task verbally as well as in writing 1 2 3 4 5

18. Tasks A and B are argumentative in nature rather than descriptive. 1 2 3 4 5

19. Task A only involves factual/concrete information. 1 2 3 4 5

20. Task B only involves factual/concrete information. 1 2 3 4 5

21. Both tasks A and B have a mixture of factual/concrete and abstract information. 1 2 3 4 5

22. The topics in the tasks are familiar to me. 1 2 3 4 5

23. The instructions for the tasks only contain words that are suitable for my level of proficiency. 1 2 3 4 5

24. The instructions for the tasks use simple, easy to understand sentence structures. 1 2 3 4 5

25. The language functions I need to perform in the task (e.g. give opinion, give reasons, provide examples etc) are clear. 1 2 3 4 5

Indicate whether you agree with the statements below about the examiner & other speakers
For task A:

26. The examiner helped me during the presentation. 1 2 3 4 5
27. I understood the examiners because they spoke at a pace that I could follow. 1 2 3 4 5
28. I find it difficult to understand an examiner who has a native speaker accent. 1 2 3 4 5

29. I am affected by the gender (M/F) of the examiners. 1 2 3 4 5
30. I prefer a female examiner. 1 2 3 4 5
31. Knowing the examiners I am talking to makes me more comfortable. 1 2 3 4 5

32. Knowing the people listening to me are my classmates makes me more comfortable. 1 2 3 4 5

33. I have no problem with another examiner being present to observe/rate my presentation. 1 2 3 4 5

For task B:

34. I was able to interact well with the other speakers/classmates because they spoke at a pace that I could follow. 1 2 3 4 5
35. I am able to interact well with the other speakers/classmates because I understand the rules of turn taking. 1 2 3 4 5
36. I find it difficult to interact with speakers/classmates who have a native speaker accent. 1 2 3 4 5
37. I am affected by the gender (M/F) of the other speakers/classmates. 1 2 3 4 5
38. I prefer to interact with same gender speakers/classmates 1 2 3 4 5
39. Knowing the speakers I am interacting with are my classmates makes me feel more comfortable. 1 2 3 4 5
40. I am happy that there is another lecturer present to observe/rate my presentation. 1 2 3 4 5

41. The following conditions in the test venue were satisfactory:
    a) Lighting 1 2 3 4 5
    b) Air-conditioning 1 2 3 4 5
    c) Background noise 1 2 3 4 5
    d) Room atmosphere 1 2 3 4 5
    e) Others, please specify

    _____ 1 2 3 4 5

    _____ 1 2 3 4 5

42. It is not possible to test everyone at the same time. There were many speaking tests conducted at different times in the last week. Indicate the degree to which you agree with the statements below.

    a) The testing environment was satisfactory and the same for every test. 1 2 3 4 5
    b) Candidates were not able to discuss exam questions between tests. 1 2 3 4 5
    c) The examiners followed specific rules on conducting the test, for e.g. that all candidates are given the same amount of time for preparation. 1 2 3 4 5
    d) The examiners followed the specifications for the test task, for e.g. all students got equally difficult questions no matter which day they took the test 1 2 3 4 5
    e) It is not possible that candidates knew the questions before the exam. 1 2 3 4 5

43. Given the conditions above for the speaking test, I would consider the possibility of doing the test using a different method where:

    a) Speech is recorded on audio tape and rated later. 1 2 3 4 5
    b) Speech is recorded on video tape and rated later. 1 2 3 4 5
    c) Test is conducted via the computer using web-based technology. 1 2 3 4 5

SECTION B: WHAT I DID IN THE TEST

For task A, rate the following statements to describe the processes you went through in order to complete the task.

Preparation stage:

1. I read the question and spent time thinking of the points I wanted to make.	1 2 3 4 5
2. I read the question and wrote down the points.	1 2 3 4 5
3. I read the question and practiced the points aloud.	1 2 3 4 5
4. I knew enough about the topic for the task from previous readings and experience.	1 2 3 4 5
5. The information in the text provided was necessary for me to complete the task.	1 2 3 4 5

Presentation stage:

6. I had prepared the speech in my mind before presenting it.	1 2 3 4 5
7. When I was presenting, I was aware of what I said and made corrections where necessary.	1 2 3 4 5
8. I was able to use words and structures that were appropriate for me to present and justify my ideas	1 2 3 4 5
9. I learnt from my speech class how to present and justify my ideas in a presentation.	1 2 3 4 5
10. I know the degree of language formality needed when responding to a lecturer's/examiner's comments or questions.	1 2 3 4 5
11. I understood the comments and questions the examiner used to help me during the presentation.	1 2 3 4 5
12. Other than the items listed above, write down any other strategies you might have used in the process of preparing and/or presenting task A.	1 2 3 4 5

_____

_____

For task B, rate the following statements to describe the processes you went through in order to complete your task.

Preparation stage:

13. We read the question together and discussed all the points together.	1 2 3 4 5
14. I read the question, then worked alone on specific aspects of the task and then we finally discussed all the points as a group.	1 2 3 4 5
15. I knew enough about the general topic of the task from previous readings and experience.	1 2 3 4 5
16. I knew enough about the specific content of the task from previous readings and experience.	1 2 3 4 5
17. The information in the instructions provided was necessary for me to complete the task.	1 2 3 4 5

Presentation stage:

18. I had prepared well in my mind what I wanted to say.	1 2 3 4 5
19. When I was presenting, I was aware of what I said and made corrections where necessary.	1 2 3 4 5

20. I was able to use words and structures that were appropriate for the task of expressing and justifying my opinion. 1 2 3 4 5
21. I was able to use words and structures that were appropriate for interacting with the group members. 1 2 3 4 5
22. I learnt from my speech class how to do all of the following: initiate, maintain and conclude a group discussion. 1 2 3 4 5
23. I understood the points made by other group members when we discussed the reasons for the solution to a problem. 1 2 3 4 5
25. I understand that in a discussion, we interact by expressing opinions, accepting and rejecting ideas/proposals, asking and giving clarifications, summarizing and concluding. 1 2 3 4 5
26. I know the degree of formality of language needed to interact with members of the group who are my classmates. 1 2 3 4 5
27. Other than the items listed above, write down any other strategies you might have used in the process of preparing and/or presenting task B.

_____

_____

SECTION C: TEACHING & LEARNING

1. Lecturers inform students of the goals, content, format and rating process of the test tasks in detail. 1 2 3 4 5
2. Lecturers spend time in class discussing with students various topics so students are familiar with information in the test. 1 2 3 4 5
3. Lecturers spend time in class practicing past year questions with students so that students are familiar with structures, vocabulary and format used in the test. 1 2 3 4 5
4. We spend a lot of time in class practicing individual speeches 1 2 3 4 5
5. We spend a lot of time in class practicing group discussions. 1 2 3 4 5
6. Because there is task A (individual presentation) in the test, I can perform better in presentations for other classes. 1 2 3 4 5
7. Because there is task B (group discussion) in the test, I can perform better in group discussions for other classes. 1 2 3 4 5
8. Because the test was designed according to the MUET syllabus, test results would interest the following parties as a measure of a candidate's speaking ability:

   a) Other higher learning institutions 1 2 3 4 5
   b) Prospective employers in the industries 1 2 3 4 5
   c) The exam syndicate of the Department of Education 1 2 3 4 5

20. I was able to use words and structures that were appropriate for the task of expressing and justifying my opinion. 1 2 3 4 5
21. I was able to use words and structures that were appropriate for interacting with the group members. 1 2 3 4 5
22. I learnt from my speech class how to do all of the following: initiate, maintain and conclude a group discussion. 1 2 3 4 5
23. I understood the points made by other group members when we discussed the reasons for the solution to a problem. 1 2 3 4 5
25. I understand that in a discussion, we interact by expressing opinions, accepting and rejecting ideas/proposals, asking and giving clarifications, summarizing and concluding. 1 2 3 4 5
26. I know the degree of formality of language needed to interact with members of the group who are my classmates. 1 2 3 4 5
27. Other than the items listed above, write down any other strategies you might have used in the process of preparing and/or presenting task B.

_____

_____

SECTION C: TEACHING & LEARNING

1. Lecturers inform students of the goals, content, format and rating process of the test tasks in detail. 1 2 3 4 5
2. Lecturers spend time in class discussing with students various topics so students are familiar with information in the test. 1 2 3 4 5
3. Lecturers spend time in class practicing past year questions with students so that students are familiar with structures, vocabulary and format used in the test. 1 2 3 4 5
4. We spend a lot of time in class practicing individual speeches 1 2 3 4 5
5. We spend a lot of time in class practicing group discussions. 1 2 3 4 5
6. Because there is task A (individual presentation) in the test, I can perform better in presentations for other classes. 1 2 3 4 5
7. Because there is task B (group discussion) in the test, I can perform better in group discussions for other classes. 1 2 3 4 5
8. Because the test was designed according to the MUET syllabus, test results would interest the following parties as a measure of a candidate's speaking ability:

   a) Other higher learning institutions 1 2 3 4 5
   b) Prospective employers in the industries 1 2 3 4 5
   c) The exam syndicate of the Department of Education 1 2 3 4 5

APPENDIX 3.2

SPEAKING TEST: STUDENT QUESTIONNAIRE

Dear student:

We are conducting research on the university speaking test. This questionnaire aims to gather data as part of this research project. We want to get your feedback on the existing speaking test.

The English Department of Pusat Bahasa conducts the speaking test every semester for students who are in the 2nd and 3rd semesters of their diploma programmes; the codes are BEL 200 and BEL 250 respectively. It is a face-to-face test where candidates work with at least two examiners. This test was designed based on the MUET (Malaysian University English Test, developed by the Malaysian Examinations Council) syllabus and it includes two major tasks of

a) an individual presentation
b) a group discussion.

The items in this questionnaire are divided according to various aspects of the test. You will thus respond to questions pertaining to the content of the test, what you did to perform the tasks and how you think the test affects teaching and learning in the English classroom.

The data collected from you will help us improve the existing test. Therefore, your cooperation in the matter is greatly appreciated.

Thank you.

Sincerely,

Saidatul A Zainal abidin
Pusat Bahasa
UiTM Shah Alam

Student number:
Gender:
Faculty:

This questionnaire covers three sections of the speaking test.: Test content, What you did during the test, and Learning and teaching in the classroom.
For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on a five point scale where:

1= Strongly disagree, 2= Disagree, 3= Undecided, 4= Agree, 5= Strongly agree
Note: DO NOT circle 3 unless you cannot understand OR really cannot answer the question.

SECTION A: TEST CONTENT

1. Task A clearly states what I am required to do in the presentation.	1 2 3 4 5
2. Task B clearly states what I am required to do in the group discussion.	1 2 3 4 5
3. Task A is a good test of my ability to communicate orally in an academic context.	1 2 3 4 5
4. Task A is a good test of my ability to speak English in social situations.	1 2 3 4 5
5. Task B is a good test of my ability to communicate orally in an academic context.	1 2 3 4 5
6. Task B is a good test of my ability to speak English in social situations.	1 2 3 4 5
7. Both task A and task B should have equal marks.	1 2 3 4 5
8. The criteria for scoring my performance (Task fulfillment, Language use, Communicative ability) were made clear to me in the instructions to the test.	1 2 3 4 5
9. The order of the tasks, i.e. task A followed by task B is appropriate for the test.	1 2 3 4 5
10. Two minutes is sufficient time for a candidate to demonstrate his/her ability to present ideas.	1 2 3 4 5
11. Ten minutes is sufficient time for a group of four candidates to demonstrate their ability to conduct a discussion on why a decision had been made.	1 2 3 4 5
12. Having written instructions to prepare for tasks A and B is helpful.	1 2 3 4 5
13. The instructions give sufficient information for candidates to prepare for tasks A and B.	1 2 3 4 5

14. Tasks A and B involve arguing for or against an idea. 1 2 3 4 5

15. Task A only contains factual/concrete information, e.g. information about physical objects, processes, people, and situations rather than abstract concepts such as love, hate, and friendship. 1 2 3 4 5

16. Task B only contains factual/concrete information, e.g. information about physical objects, processes, events, and people. 1 2 3 4 5

17. Both tasks A and B have a mixture of factual/concrete and abstract information. 1 2 3 4 5

18. The topic in tasks A and B is familiar to me. 1 2 3 4 5

19. The instructions for the tasks only contain words that are suitable for my level of language ability. 1 2 3 4 5

20. The instructions for the tasks use simple, easy to understand sentence structures. 1 2 3 4 5

21. The language functions I need to perform in the task (e.g. give opinion, give reasons, provide examples etc) are clear. 1 2 3 4 5

Indicate whether you agree with the statements below about the examiner & other speakers
For task A:

22. The examiner gave us extra help during the presentation. 1 2 3 4 5
23. I understood the examiners because they spoke at a pace that I could follow. 1 2 3 4 5
24. I find it difficult to understand an examiner because of his/her accent. 1 2 3 4 5
25. I am happy with either a male or a female examiner. 1 2 3 4 5
26. I prefer the same gender examiner. 1 2 3 4 5
27. Knowing the examiners I am talking to makes me more comfortable. 1 2 3 4 5
28. Knowing the people listening to me are my classmates makes me more comfortable. 1 2 3 4 5

29. I am happy with another examiner being present to observe/mark my language use in task A. 1 2 3 4 5

For task B:

30. I was able to interact well with the other speakers because they spoke at a speed that I could follow. 1 2 3 4 5
31. I am able to interact well with the other speakers because I understand the rules of turn taking. 1 2 3 4 5
32. I find it difficult to interact with speakers in task B because of their accent. 1 2 3 4 5
33. I am happy working with a male or female speaker. 1 2 3 4 5
34. I prefer to interact with the same gender speakers. 1 2 3 4 5
35. Knowing the speakers I am interacting with are my classmates makes me feel

more comfortable. 1 2 3 4 5

36. I am happy that there is another examiner present to observe/mark my language use in task B. 1 2 3 4 5

37. The following conditions in the test venue were OK:

a) Lighting 1 2 3 4 5
b) Noise level 1 2 3 4 5
c) Room temperature 1 2 3 4 5
d) Seating arrangement 1 2 3 4 5
e) Conditions for disabled students 1 2 3 4 5

38. It is not possible to test everyone at the same time. There were many speaking tests conducted at different times in the last week. Please tell us the degree to which you agree with the statements below by circling the number **that REFLECTS YOUR VIEWPOINT**

a) Candidates were not able to discuss exam questions with students who had already taken the test. 1 2 3 4 5
b) The examiners followed the rules specific for conducting the test, for e.g. that all candidates are given the same amount of time for preparation. 1 2 3 4 5

39. The following would be more helpful ways of preparing for Tasks A and B:

a) Listen to a recorded presentation (lecture, dialogue, announcement, etc.) 1 2 3 4 5
b) Watch a video presentation 1 2 3 4 5
c) Look at visual text (pictures, graphs, maps, brochures, etc.) 1 2 3 4 5
d) Hearing the instructions read out to you as well as seeing them in writing 1 2 3 4 5

40. Given the conditions above for the speaking test, I would consider the possibility of doing the test using a different method where the test is conducted by the computer. 1 2 3 4 5
(In this test, everything is the same as in the face-to-face test; the format, topic, content, marking criteria, and so on. The only difference is in the method of delivery, i.e. via the computer)

SECTION B: WHAT I DID IN THE TEST

This section concerns the processes you went through in order to complete both task A & task B.

For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on the five point scale.

1= Strongly disagree 2= Disagree 3= Undecided 4= Agree 5= Strongly agree

Note: **DO NOT circle 3 unless you cannot understand OR really cannot answer the question.**

TASK A

Preparation stage:

1. I read the task very carefully to understand what was required.	1 2 3 4 5
2. I thought of the points I wanted to make.	1 2 3 4 5
3. I thought of how to satisfy the examiners.	1 2 3 4 5
4. I wrote down the points I wanted to make.	1 2 3 4 5
5. I thought of the words and expressions I needed to fulfill the task.	1 2 3 4 5
6. I thought of the structures I needed to fulfill the task.	1 2 3 4 5
7. I practised the speech in my mind.	1 2 3 4 5
8. I was familiar with the general topic of the task from previous readings and experience.	1 2 3 4 5
9. I knew enough specific information for the task from previous readings and and experience.	1 2 3 4 5
10. The information in the instructions provided was necessary for me to complete the task.	1 2 3 4 5

Presentation stage

11. During my presentation, I checked:

a) the appropriateness of the words that I used	1 2 3 4 5
b) my grammatical accuracy	1 2 3 4 5
c) the organization of my presentation	1 2 3 4 5

12. During my presentation, I adjusted:

a) the appropriateness of the words that I used	1 2 3 4 5
b) my grammatical accuracy	1 2 3 4 5
c) the organization of my presentation	1 2 3 4 5

13. I know I have to talk differently to a lecturer than to my friends in the class. 1 2 3 4 5

TASK B

Preparation stage:

14. I read the task very carefully to understand what was required.	1 2 3 4 5
15. I thought of the points I wanted to make.	1 2 3 4 5

16. I thought of how to satisfy the examiners. 1 2 3 4 5
17. I wrote down the points I wanted to make. 1 2 3 4 5
18. I thought of the words and expressions I needed to fulfill the task. 1 2 3 4 5
19. I thought of the structures I needed to fulfill the task. 1 2 3 4 5
20. The information provided by other students in task A helped me complete this task. 1 2 3 4 5
21. The information in the instructions was necessary for me to complete the task. 1 2 3 4 5

Discussion stage:

24. When I spoke during the discussion, I checked:

a) the appropriateness of the words that I used 1 2 3 4 5
b) my grammatical accuracy 1 2 3 4 5
c) the effect of what I said on other speakers 1 2 3 4 5

25. When I spoke during the discussion, I adjusted:

a) the appropriateness of the words that I used 1 2 3 4 5
b) my grammatical accuracy 1 2 3 4 5
c) the point(s) I wanted to make 1 2 3 4 5

26. When others are speaking, I checked the points they made. 1 2 3 4 5

27. Based on what they said, I adjusted my next response. 1 2 3 4 5

28. In the test, I had no problem doing the following:

a) initiating a discussion 1 2 3 4 5
b) keeping a conversation going 1 2 3 4 5
c) connecting what I said to what has just been said 1 2 3 4 5
d) taking my turn appropriately 1 2 3 4 5
e) concluding a group discussion 1 2 3 4 5

29. When talking in English to my classmates, I use different language than when I talk to my lecturers. 1 2 3 4 5

30. We were able to conduct the discussion smoothly. 1 2 3 4 5

31. We were able to conduct the discussion in an organized fashion. 1 2 3 4 5

SECTION C: TEACHING & LEARNING

This section concerns the effect the test has on teaching and learning in the English classroom. For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on the five point scale.

1= Strongly disagree 2= Disagree 3= Undecided 4= Agree 5= Strongly agree

Note: **DO NOT circle 3 unless you cannot understand OR really cannot answer the question.**

1. Lecturers inform students of the goals, content, format and rating process of the test tasks in detail.	1 2 3 4 5
2. Lecturers spend time in class discussing with students various topics so students are familiar with information in the test.	1 2 3 4 5
3. Lecturers spend time in class practicing past year questions with students so that students are familiar with structures, vocabulary and format used in the test.	1 2 3 4 5
4. We spend a lot of time in class practicing individual speeches	1 2 3 4 5
5. We spend a lot of time in class practicing group discussions.	1 2 3 4 5
6. I learnt from my speech class how to present and support my ideas in a presentation.	1 2 3 4 5
7. I learnt from my speech class how to:	
a) initiate a discussion	1 2 3 4 5
b) keep a conversation going	1 2 3 4 5
c) connect what I said to what has just been said	1 2 3 4 5
d) take my turn appropriately	1 2 3 4 5
e) conclude a group discussion	
8. Because I have had practice in individual presentations (task A) for the test, I am able to perform in presentations in other classes.	1 2 3 4 5
9. Because I have had practice in group discussions (task B) for the test, I am able to participate in group discussions in other classes.	1 2 3 4 5

APPENDIX 3.3

STAFF SPEAKING TEST QUESTIONNAIRE

Dear Respondent:

We are conducting research on the university speaking test. This questionnaire aims to gather data as part of this research project. We want to get your feedback on the existing speaking test.

The English Department of Pusat Bahasa conducts the speaking test every semester for students who are in the 2nd and 3rd semesters of their diploma programmes; the codes are BEL 200 and BEL 250 respectively. It is a face-to-face test where candidates work with at least two examiners. This test was designed based on the MUET (Malaysian University English Test, developed by the Malaysian Examinations Council) syllabus and it includes two major tasks of

a) an individual presentation
b) a group discussion.

The items in this questionnaire are divided according to various aspects of the test such as the content of the test, what students did to perform the tasks, how the test affects teaching and learning in the English classroom, scoring validity, and criterion-related validity.

Your feedback will be treated as confidential and will be used for the purpose of this research only.

We hope that the data collected from you will help us improve the existing test.

Therefore, your cooperation in the matter is greatly appreciated.

Thank you.

Sincerely,

Saidatul A Zainal abidin
Lecturer of English
Pusat Bahasa
UiTM Shah Alam

This questionnaire covers five aspects of the speaking test: Test content, What students did during the test, The impact on teaching and learning, Scoring validity, and Criterion-related validity.
For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on a five point scale where:

1= Strongly disagree, 2= Disagree, 3= Undecided, 4= Agree, 5= Strongly agree
Note: DO NOT circle 3 unless you cannot understand OR really cannot answer the question.

SECTION A: TEST CONTENT

1. Task A clearly states what candidates are required to do in the presentation.	1 2 3 4 5
2. Task B clearly states what candidates are required to do in the group discussion.	1 2 3 4 5
3. Task A is a good test of a candidate's ability to communicate orally in an academic context.	1 2 3 4 5
4. Task A is a good test of a candidate's ability to speak English in social situations.	1 2 3 4 5
5. Task B is a good test of a candidate's ability to communicate orally in an academic context.	1 2 3 4 5
6. Task B is a good test of a candidate's ability to speak English in social situations.	1 2 3 4 5
7. Both task A and task B should have equal marks.	1 2 3 4 5
8. The criteria for scoring the performance (Task fulfillment, Language use, Communicative ability) were made clear to candidates in the instructions to the test.	1 2 3 4 5
9. The order of the tasks, i.e. task A followed by task B, is appropriate for the test.	1 2 3 4 5
10. Two minutes is sufficient time for a candidate to demonstrate his/her ability to present ideas.	1 2 3 4 5
11. Ten minutes is sufficient time for a group of four candidates to demonstrate their ability to conduct a discussion on why a decision had been made.	1 2 3 4 5
12. Having written instructions to prepare for tasks A and B is helpful.	1 2 3 4 5
13. The instructions give sufficient information for candidates to prepare for tasks A and B.	1 2 3 4 5
14. Tasks A and B involve arguing for or against an idea.	1 2 3 4 5

15. Task A only contains factual/concrete information, e.g. information about physical objects, processes, people, and situations, rather than abstract concepts such as love, hate, and friendship. 1 2 3 4 5

16. Task B only contains factual/concrete information, e.g. information about physical objects, processes, events, and people, rather than abstract concepts such as love, hate, and friendship. 1 2 3 4 5

17. Both tasks A and B have a mixture of factual/concrete and abstract information. 1 2 3 4 5

18. The topic in tasks A and B is familiar to candidates. 1 2 3 4 5

19. The instructions for the tasks only contain words that are suitable for a candidate's level of language ability. 1 2 3 4 5

20. The instructions for the tasks use simple, easy to understand sentence structures. 1 2 3 4 5

21. The language functions that candidates need to perform in the task (e.g. give opinion, give reasons, provide examples etc) are clear. 1 2 3 4 5

Indicate whether you agree with the statements below about the examiner & other speakers

For task A:

22. The examiner gives extra help during the presentation. 1 2 3 4 5

23. A candidate would understand the examiners because they spoke at a pace that he/she could follow. 1 2 3 4 5

24. Candidates find it difficult to understand an examiner because of his/her accent. 1 2 3 4 5

25. A candidate is happy with either a male or a female examiner. 1 2 3 4 5

26. Candidates prefer the same gender examiner. 1 2 3 4 5

27. Candidates are more comfortable if they know the examiners they are talking to. 1 2 3 4 5

28. Candidates are more comfortable if they know the people who are listening to them are their classmates. 1 2 3 4 5

29. Candidates are happy with another examiner being present to observe/mark their language use in task A. 1 2 3 4 5

For task B:

29. A candidate is able to interact well with the other speakers because they spoke at a speed that he/she could follow. 1 2 3 4 5

30. A candidate is able to interact well with the other speakers because he/she understands the rules of turn taking. 1 2 3 4 5

32. A candidate would find it difficult to interact with speakers in task B because of their accent. 1 2 3 4 5

33. Candidates are happy working with a male or female speaker. 1 2 3 4 5

34. Candidates prefer to interact with the same gender speakers. 1 2 3 4 5

35. Candidates are more comfortable if they know the speakers they are interacting with are their classmates. 1 2 3 4 5

36. Candidates are happy with another examiner being present to observe/mark their language use in task B. 1 2 3 4 5

37. The following conditions in the test venue are satisfactory:

a) Lighting	1 2 3 4 5
b) Noise level	1 2 3 4 5
c) Room temperature	1 2 3 4 5
d) Seating arrangement	1 2 3 4 5
e) Conditions for disabled students	1 2 3 4 5

38. Because it is not possible to test everyone at the same time, many speaking tests are conducted at different times within a week. Please tell us the degree to which you agree with the statements below by circling the number **that REFLECTS YOUR VIEWPOINT**

a) Candidates are not able to discuss exam questions with others who had already taken the test.	1 2 3 4 5
b) The examiners followed rules that are specified for conducting the test, for e.g. that all candidates are given the same amount of time for preparation.	1 2 3 4 5

39. The following would be more helpful ways of preparing for Tasks A and B:

a) Listen to a recorded presentation (lecture, dialogue, announcement, etc.)	1 2 3 4 5
b) Watch a video presentation	1 2 3 4 5
c) Look at visual text (pictures, graphs, maps, brochures, etc.)	1 2 3 4 5
d) Hearing the instructions read out as well as seeing them in writing	1 2 3 4 5

40. Given the conditions above for the speaking test, I would consider the possibility of the test being conducted by the computer. (In this test, everything is the same as in the face-to-face test; the format, topic, content, marking criteria, and so on. The only difference is in the method of delivery, i.e. via the computer) 1 2 3 4 5

SECTION B: WHAT STUDENTS DO IN THE TEST

This section concerns the processes a student goes through in order to complete the task; it has to do with cognitive processing. Items in this section reflect YOUR viewpoint/perspective of what students would do, and how they would attempt the task, in order to complete it.

TASK A: For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on the five point scale.

1= Strongly disagree 2= Disagree 3= Undecided 4= Agree 5= Strongly agree
Note: DO NOT circle 3 unless you cannot understand OR really cannot answer the question.

Preparation stage:

1. Students read the task very carefully to understand what is required. 1 2 3 4 5
2. Students think of the points they want to make. 1 2 3 4 5
3. Students think of how to satisfy the examiners. 1 2 3 4 5
4. Students write down the points they want to make. 1 2 3 4 5
5. Students think of the words and expressions they need to fulfill the task. 1 2 3 4 5
6. Students think of the structures they need to fulfill the task. 1 2 3 4 5
7. Students practise the speech in their minds. 1 2 3 4 5
8. Students are familiar with the general topic of the task from previous readings and experience. 1 2 3 4 5
9. Students know enough specific information for the task from previous readings and and experience. 1 2 3 4 5
10. The information in the instructions provided is necessary for students to complete the task. 1 2 3 4 5

Presentation stage

11. During the presentation, students check:

    a) the appropriateness of the words that they use 1 2 3 4 5
    b) the accuracy of their grammar 1 2 3 4 5
    c) the organization of their presentation 1 2 3 4 5

12. During the presentation, students adjust:

    a) the appropriateness of the words that they use 1 2 3 4 5
    b) the accuracy of their grammar 1 2 3 4 5
    c) the organization of their presentation 1 2 3 4 5

13. Students know they have to talk differently to a lecturer than to their friends in the class. 1 2 3 4 5

TASK B : For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on the five point scale.

1= Strongly disagree 2= Disagree 3= Undecided 4= Agree 5= Strongly agree

Preparation stage:

14. The information provided by other students in task A can help a student to complete this task.	1 2 3 4 5
15. The information in the instructions is necessary for students to complete the task.	1 2 3 4 5
16. Students read the task very carefully to understand what is required.	1 2 3 4 5
17. Students think of the points they want to make.	1 2 3 4 5
18. Students think of how to satisfy the examiners.	1 2 3 4 5
19. Students write down the points they want to make.	1 2 3 4 5
20. Students think of the words and expressions they need to fulfill the task.	1 2 3 4 5
21. Students think of the structures they need to fulfill the task.	1 2 3 4 5

Discussion stage:

22. When students speak during the discussion, they check:

a) the appropriateness of the words that they use	1 2 3 4 5
b) the accuracy of their grammar	1 2 3 4 5
c) the effect of what they say on other speakers	1 2 3 4 5

23. When students speak during the discussion, they adjust:

a) the appropriateness of the words that they use	1 2 3 4 5
b) the accuracy of their grammar	1 2 3 4 5
c) the point(s) they want to make	1 2 3 4 5

24. When others are speaking, a student would check the points they make.	1 2 3 4 5
25. Based on what they said, the student would adjust his/her next response.	1 2 3 4 5

26. In the test, students had no problem doing the following:

a) initiating a discussion	1 2 3 4 5
b) keeping a conversation going	1 2 3 4 5
c) connecting what is said to what has just been said	1 2 3 4 5
d) taking turns appropriately	1 2 3 4 5
e) concluding a group discussion	1 2 3 4 5

27. When talking in English to their classmates, students use different language than when they talk to lecturers.	1 2 3 4 5
28. Students are able to conduct the discussion smoothly.	1 2 3 4 5
29. Students are able to conduct the discussion in an organized fashion.	1 2 3 4 5

SECTION C: TEACHING & LEARNING

This section concerns the effect the test has on teaching and learning in the English classroom. For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on the five point scale.

1= Strongly disagree 2= Disagree 3= Undecided 4= Agree 5= Strongly agree

Note: DO NOT circle 3 unless you cannot understand OR really cannot answer the question.

1. Lecturers give students full details of all aspects of the tests tasks (e.g. goals, content, format and rating process) 1 2 3 4 5

2. Lecturers spend time in class discussing with students various topics so students are familiar with information required in the test. 1 2 3 4 5

3. Lecturers spend time in class practicing past year questions with students so that students are familiar with structures, vocabulary and format used in the test. 1 2 3 4 5

4. Students spend a lot of time in class practicing individual speeches 1 2 3 4 5

5. Students spend a lot of time in class practicing group discussions. 1 2 3 4 5

6. Students learn from their speech class how to present and support ideas in a presentation. 1 2 3 4 5

7. Students learn from their speech class how to:
   a) initiate a discussion 1 2 3 4 5
   b) keep a conversation going 1 2 3 4 5
   c) connect what they say to what has just been said 1 2 3 4 5
   d) take their turn appropriately 1 2 3 4 5
   e) conclude a group discussion 1 2 3 4 5

9. Because students have had practice in individual presentations (task A) for the test, they are able to perform better in presentations in other classes. 1 2 3 4 5

10. Because students have had practice in group discussions (task B) for the test, they are able to participate better in group discussions in other classes. 1 2 3 4 5

SECTION D: SCORING VALIDITY

Both task A & task B have the same rating scale. This rating scale consists of three components: Task fulfillment, Language use, Communicative ability.

For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on the five point scale.

1= Strongly disagree 2= Disagree 3= Undecided 4= Agree 5= Strongly agree

Note: DO NOT circle 3 unless you cannot understand OR really cannot answer the question.

a) The components Task fulfillment, Language use, and Communicative ability cover all aspects of the performance that the examiner looks for in the presentation.
1 2 3 4 5

b) Three components are enough for markers to use in making a fair judgment of the oral tasks.
1 2 3 4 5

c) The criteria for rating are clear to all markers.
1 2 3 4 5

d) The raters have been given enough information on the procedures for rating the tasks appropriately.
1 2 3 4 5

e) The raters are well trained in using all the rating procedures for the test. 1 2 3 4 5

f) The raters are standardized to benchmark candidate performance levels before marking/rating begins.
1 2 3 4 5

g) The raters are able to work without any disturbance and distraction during the rating process.
1 2 3 4 5

h) Raters' marks are moderated after the test to sort out any differences or problems in the marking.
1 2 3 4 5

i) Statistical analyses are conducted on the marks to check consistency and level of rating.
1 2 3 4 5

j) There are two lecturers present at the test; only the interlocutor interacts with the candidate.
1 2 3 4 5

k) Before the results are issued to candidates, the exam committee checks all results to ensure fairness, especially for those who are close to the pass/fail boundary.
1 2 3 4 5

SECTION E: CRITERION-RELATED VALIDITY

A test is said to have criterion-related validity if a relationship can be demonstrated between test scores and some external criterion which is believed to be a measure of the same ability.

For each of the items below, circle the number that REFLECTS YOUR VIEWPOINT on the five point scale.

1= Strongly disagree 2= Disagree 3= Undecided 4= Agree 5= Strongly agree

Note: DO NOT circle 3 unless you cannot understand OR really cannot answer the question.

a) There have been attempts to compare test scores from this test with other measures of candidates' language ability such as teacher assessments or continuous assessments.
1 2 3 4 5

b) There have been attempts to compare test scores from this test with other external measures such as other tests of spoken language.
1 2 3 4 5

c) Candidates' test scores have been compared to scores from other similar oral tests conducted at the faculty by subject lecturers.
1 2 3 4 5

d) Candidates' test scores have been compared to results of self-assessment of their own language abilities.
1 2 3 4 5

THANK YOU.
SAZA
12/11/03

APPENDIX 3.4

STUDENT Interview Notes/Summary

INTERVIEW DATA COLLECTED IN MAIN STUDY 1: JAN 26 – MAR'09 2004

Notes:

- Interviews were conducted as soon as the speaking test began on 26/01/04 in the main campus in Shah Alam as well as all branch campuses in West and East Malaysia. Campuses involved in all interviews: main campus Shah Alam, Jengka campus in the state of Pahang, Segamat campus in the state of Johor, and Kota Kinabalu campus in the state of Sabah, East Malaysia.
- Student interviews were conducted in groups of four as the test was conducted in groups of four students. Total number interviewed = 16 groups from 9 faculties (EE, CE, ME, AP, AC, ASc, CS, BM, AL)
- Examiner interviews were conducted in pairs as they would have conducted the test together. Total number interviewed = 7
- Administrator and expert interviews were conducted individually. Total number interviewed: Administrator = 4; Expert = 2
- Refer to the 'Roster' and 'Schedule of Main Study 1' for other details on the nature of data collection for Interviews
- All interviews were audio-tape recorded
- All interview data (students and other participants) are analyzed by
  a) listening to recordings and noting points that have been highlighted (ref table below)
  b) entering transcriptions of the interview in Microsoft Word
  c) using Hyper Research software for qualitative data analysis

INTERVIEWS WITH STUDENTS

Date/Faculty/ Participant/ Campus	Stage I Preparation: Task A/B	Factors	Stage II Presentation: Task A/B	Factors	Stage III Fair? Differences? Plan? Monitor? Other factors?
Jan26 *Shah Alam* *EE* Md Hafizal Md Azri Md Sanusi Mohamed	A: Brief notes – main pts/ examples B: Wrote pts from presn A	•Purpose not immediately clear •2 min too short for prep •Acquaintance-ship impt	A: Couldn't present all points/read, elaborate some B: Fair discussion	•Time to prepare •Nerves •Knowledge on topic •Language ability •Impt to know other speakers	-Fair -A more difficult than B -Not enough time to plan -Didn't monitor **Other factors:** * Class practice/prep helped * Don't read enough
Shah Alam *CE* A Firdaus A Osman Zainizam Md Affendi	A: Wrote brief points B: Go through all points fr A	•2 min not enough to prepare A •Lost time looking for test room •Topic familiar but difficult •Impt to know other speakers	A: Presented what's written B: Problems when there are disagreements	•2 min not enough to prepare •Topic familiar but difficult •Impt to know other speakers	- Fair, with practice - No real difference between A&B - Speech quite spontaneous - Didn't monitor but conscious of others' errors **Other factors:** *Language ability *Time of test; late in the day; not a good time *Topic difficulty

Jan27 ***Shah Alam AP*** Faizal - Jubley N Razi	A: Read, list points, elaborate B: Prepared based on points from A	•Format- Indv related to group discussion •2 min not sufficient to prepare well •Topic quite difficult •Acquaintance-ship impt/ mixed group didn't help ••Format of question paper: wld help if Task A & B were on separate pages •Fluency more impt than accent	A: Presented what's been written; not enough time to organize pts well Elaborate pts during presn B: Listening to other speakers very impt for discussion	•Time real pressure •Class practice helped; knowing what to expect •Instructions quite clear	- A not so fair - Difference in task format but content in A related to B - Insufficient time to plan - Monitoring impt to perform well, esp in B; own spch & others' **Other factors:** *Focus on own spch & others' esp content *Candidacy *Test encourages usage of English
Shah Alam *ME* Nazri Iliana Mohamed	A: List down many pts/ sentences B: Listen to A to make notes/mental preparation	•Purpose: candidacy dep •Time insufficient to think thru pts; candidacy dep •Acquaintance-ship impt	A: During presentation worried abt *spch rate/ time *use of words B: Speaking with/to others better; help/support from others	•Listening carefully to A/A affects B •Other speakers' spch rate •Preparation time	-A not so fair: dependant on *candidacy * topic you get - B easier to present than A; ideas came spontaneously - Insufficient time to plan **Other factors:** * Candidacy; pressure on candidate A * Lack of practice * Language ability

Jan28 ***Shah Alam AC*** Mariaty Norhaida Nisha Nor	A: Wrote points/elaborate some B: Consider pts made in A *Consider examiners	•Topic familiar but lack background knowledge •Knowing A is related to B helps •Little knowledge of rating criteria •Time too short/knowing this puts pressure on candidate A •Ph cond- Rain was distracting	A: Presented points prepared B:Concentrated on the discussion	•Candidacy •Time • Linguistic features (nature of information, language) not so clear	- Not fair esp task A - Differences between presentations NOT clear/gd pts no elab, one pt elab a lot (?) - Planned but during presn, sometimes spontaneous **Other factors**: * Knowing some factors helped: time, format, topics
Shah Alam AC Jalilah Nabilah Norlizazilah Zalina	A: Wrote brief notes/points B: Prepare individually but discussed on who does what	•Purpose quite confusing (2 tasks in one qn?) •Enough time just to list points/cand A disadvantaged •Other examiner not familiar	A: Spontaneous elaboration of points B: Discussion went quite well	•Speaking to & with others made us more relaxed	- A not fair/2 minutes too short; task B more realistic - No major difference/A affects B - Planned; ideas, structure, even vocab, but couldn't present all - Monitored as discussion proceeded/ listened to others **Other factors**: * Topic familiarity/not discussed out of class a lot *Candidacy esp cand A/ Others have more time to think/plan

Jan 29 ***Shah Alam*** *ASc* Nurharniada Nurbaizura Nurmiza	A: Wrote brief notes B: Wrote what's presented in A/already decided each other's role in discussion	•Time too short for preparation • Topic familiar, had some dis in class	A: Presented what's written B: Discussion didn't go too well/difficult when there's disagreement, to conclude	• Time effect: wanted to end speech in A; have to spk a lot in B • Candidacy where points are concerned/ point given to present	- Unfair, esp if get difficult 'point' to present - A more difficult than B - Planned how to present in the mind -No monitoring **Other factors**: * Language ability * Topic diffy dependant on candidacy
Shah Alam *ASc* Saidatul Siti Hajjar Siti Hajar Salfariza	A: Wrote main points B: Discussed quickly which option to choose~ planned during prep time which point to agree on & why…	•Ph cond – Noise from construction work nearby! •Time – 2 min insufficient to prepare well	A: Think abt how to elaborate pts/ spontaneous/ left out some pts B: More involved in the discussion	•Format of grp discussion encouraged us to spk more	- Fair if 'point' given to present quite easy - B has more involvement frm everyone - Planned how to conduct discussion - Monitor the content, esp in task B **Other factors**: * Test admin; waiting outside for turn * Lack b'ground know on topic * Language ability v impt

Feb 03 *Shah Alam* *CS* Adam Cand B Cand C Husni	A: Wrote brief notes + some *mental* preparation B: Wrote what's presented in A/decided each other's role quickly (who leads, who concludes, etc)	•Time for preparation too short • Response format impt : A influences B • Test length impt: knowing this, you time yourself for prep & presn • Knowing other speakers impt: know how to interact with them • Topic familiar, had some dis in class but insufficient content know.	A: Presented what's written but worried abt score/repetition of words& ideas B: Discussion was fair/ difficult when others have the same ideas…	• Time effect: have to do a lot: points, elaboration, how to deliver • Rating criteria: don't know details, wld help if we knew • Question clear --- not so When faced with difficulty, examiner helps	- Fair but A quite challenging - Difference: A focus on one point; B others to consider - Planned: Words to use From Malay to English Organize pts in order - Some monitoring, esp repeating words/points **Other factors**: * Examiner 2: not familiar * Time of test: AM better than PM * Examiner's mannerism, attitude etc. *Positive-preparation for MUET!

Feb 03 UiTM KK, Sabah BM Analisa S Normalia Norlizah S Nurain	A: Wrote brief notes B: Went through pts in A	▪ Presence of others esp examiner 2 ▪ Time too short ▪Thought of low marks/ failing ▪ Topic quite difficult	A: Presented what's written/ problem expressing points well B: Discussed through, then concluded	▪ Language difficulty in elaborating points/some spontaneity ▪Translating from BM to English	- Fair because lots of discussion on topics/ task B helps if you had prob in A - Task B easier since roles planned already - Planned for task B - Monitored each others sph in task B **Other factors**: * Lack b'ground know on the topic
BM Caroline Natalie Laura Edna	A: Read, u'stand, wrote main pts B: Listen to presentations in A/mental preparation	▪Time for prep too short ▪ Nerves/Test cond: knowing this is graded! ▪Setting/Sitting position; examiners intimidating	A: Presented what's written/ candidate A dis-advantaged B: Conducted discussion well: listen to others	▪Topic very familiar! – (brainstormed 22 topics in class) ~ so much discussion in class on topics: in test, different topics	- B is fair/A dependant on candidacy - Different: A focus on yourself; B focus on others - Planned as presentations went on - Monitor each other's speech, esp content – correct own vocabulary; my friends corrected me **Other factors**: * Distractions as test was conducted in language lab; also noise from outside

Feb 04 UiTM KK, Sabah DPA Dg Razlina Shila - Sylvester Diana – Iskandar Adriana – Amaludin	A: Time constraint so more *mental* preparation B: Points here related to A – *mental* preparation	•Purpose not too clear; seek help from examiner • Topic familiar but insufficient b'ground know • 2 min not enough to prepare well •Very impt that we know other speakers & examiners	A: Mix of what's written & spontaneous speech when elaborating pts/lots of repetition too B: Not a good discussion – others come in when you're loss for words; some interrupt & spoil flow of discussion	• Time for presentation a little short • Anxious when everyone wanted to say s'thing; some interject instead (!) •Knowing other speakers well also makes it personal (!)	- B quite unfair because give a chance for gd spkr to dominate! - Different: individual vs teamwork - Planned esp for content - Monitored own spch & content; irrelevant pts rejected **Other factors**: * Time management in task B * Organization in task B impt; can be chaotic
DPA Finella Harisan	A: Wrote brief points B: Based on what others said in task A – *mental* notes	• 2 min not enough to prepare well • Topic familiar but more impt to listen to points assigned to each candidate	A: Presented what's written & some ideas were spontaneous B: Discussion cld be better if listen carefully to others	• Time too short to present esp. A •Knowing other speakers helped	- Fairness related to topic/idea assigned to ea candidate - Monitored language use a bit - B slightly easier than A bec of interaction with others **Other factors:** * Nerves esp in task A; focus all on you * Presence of others in the room

Feb 06 Shah Alam AP N Razi Maizura Md Rodzahn	A: Wrote points/organize them B: More points & elaborations, based on task A	▪ Format of qns: Would be easier to read if each task is on individual page ▪ Time for prep too short ▪ Knowing other speakers helps: problem being mixed with other groups ▪ Topic familiar discussed in class & media	A: Presented most of the points B: Other speakers affect flow of discussion/ listening to others v impt!	▪ NOT knowing other speakers is a problem ▪ Knowing examiner is class lecturer helps a lot ▪ Task B: have to listen carefully to others	- Fair but problems with: Prep time Format of qn/ reading it Test setting: too "formal" - Difference is B easier to prepare: A related to B - Planned how to organize & present ideas - Monitored occasionally; esp. content/ conscious of language prob **Other factors**: * Knowledge of test from BEL200 helps * B'ground know of topic * Test anxiety * Preparations (in & out of class) before test v impt * Environment i.e. test setting * Test admin i.e. examiner keeping time, mannerisms, etc.

Feb 10 UiTM Jengka, Pahang BM A Zakuan Md Rafiq Zawati Noorul Huda	A: Wrote points, thought of how to elaborate them B: Same as A; more points to think of	• Purpose not too clear • Test conducted in lecturer's room, not too conducive • Prep time too short • Rating criteria: not well informed •Topic familiar but points assigned to candidates NOT balanced; some more difficult than others • Knowing other speakers v impt	A: Elaborated points listed down/ show confidence through body language! B: Discussion v dependant on other speakers; if they are quiet OR agree too quickly - difficult to proceed	• Body language/ gestures during presentation • Knowing other speakers & examiners • Purpose of task: v impt that it's clear • Other speakers: spch style, language, etc.	- Not so fair : v topic/candidacy dependant (A, B, C, D) - fairly difficult - Not different: A related to/affects B - Planned how to score!: listen well to others' pts; agree/disag strength/weak-ness - Monitor content > lang; impt to get points across **Other factors:** * Difficulty level of ideas assigned to candidates within topic * Language ability v impt

Feb 11 UiTM Segamat, Johor BM Farha Ismail Nor Rashidah Norhayati Hazah	A: Thought of points & "mapped" them – select word/phrases to use B: Based on presn in A, chose one with most advantages	▪ Purpose not too clear ▪ Rating criteria: how you do in A affects B ▪ Test setting caused high anxiety ▪ Time to prepare too short ▪ Topic not too familiar: insufficient content know & experience (Raising funds for orphanage.... Medical camp....) ▪ 2nd examiner: the way he stared & didn't help	A: Presented what's written/ also according to "map" B: Listen to others and speak; the only way, quite spontaneous	▪ Knowing other speakers helps a lot ▪ Examiners' mannerism, voice ▪ Lack of practice in class, esp task A	- Fair test but v dependant on topic - B different from A: less anxious & other speakers help/support you - Planned before test, not enough time during it; B went better than A - Conscious of lang errors but couldn't monitor/ examiner indicated them! **Other factors:** * Test setting/atmosphere * Content knowledge * Language ability v impt * Practice, in class & outside: language use & reading up

APPENDIX 3.5

STAFF Interview Notes/Summary

INTERVIEW DATA COLLECTED IN MAIN STUDY 1: JAN 26 - MAR'09 2004

Notes:

- Interviews were conducted as soon as the speaking test began on 26/01/04 in the main campus in Shah Alam as well as all branch campuses in West and East Malaysia. Campuses involved in all interviews: main campus Shah Alam, Jengka campus in the state of Pahang, Segamat campus in the state of Johor, and Kota Kinabalu campus in the state of Sabah, East Malaysia.
- Student interviews were conducted in groups of four as the test was conducted in groups of four students. Total number interviewed = 16 groups from 9 faculties (EE, CE, ME, AP, AC, ASc, CS, BM, AL)
- Examiner interviews were conducted in pairs as they would have conducted the test together. Total number interviewed = 7
- Administrator and expert interviews were conducted individually. Total number interviewed: Administrator = 4; Expert = 2
- Refer to the 'Roster' and 'Schedule of Main Study 1' for other details on the nature of data collection for Interviews
- All interviews were audio-tape recorded
- All interview data (students and other participants) are analyzed by
  a) listening to recordings and noting points that have been highlighted (ref table below)
  b) entering transcriptions of the interview in Microsoft Word
  c) using Hyper Research software for qualitative data analysis

INTERVIEWS WITH OTHER PARTICIPANTS

Date/ Participant/ Campus	Context validity	Theory-based validity	Scoring validity	Consequential validity	Criterion-related validity	Other factors/ comments
Feb-04 *UiTM KK, Sabah* *Lecturer/ Examiner* *Geeta S* *Jasman J*	TASK * Purpose clear *Prep time: too short esp for wk std *Rating- std not informed in detail Criteria unclear; vague descriptions - usually 2 examiners- discrepancy in marking bet raters SETTING *Venue at office, lab... minor distractions *Uniformity: diff to control, instructions fr SA, pair junior with senior raters *Security Students tell each other topics/2 sets allowed per day DEMANDS *Language used appropriate Instructions clear	*Language a major prob; busy *translating* Knowledge is lacking; exposure dep on whether they come fr villages or towns *Processing: NOT able to analyse, think critically, ideas not well thought of or argued * Preparation Plan too quickly, look at key words & phrases; don't fulfil requirements of task	*Examiners meet and moderate scores; ref Coordinator * Analysis done on final scores, not spk independently *Training: 1.workshop on MUET grading 2. junior lecturers paired with senior lecturers	*Dependent on constraint of semester; time a major problem – pressure on lecturers to prepare them for test	* Don't really compare the scores across the skills: L/S/R/W *Occasionally, some score high in spk but lower in writing	Factors for test developers in SA to consider: *EM students are different from WM students in language ability & exposure to current issues- task demands *Opportunities for training of raters from this campus *Lecturer input into final marks for speaking ~ 5%

	*Occasionally, topics not familiar for rural EM std *Examiner helps with difficult words *Between topics difficulty level not balanced; some more diff than others					
Feb 10 *UiTM Jengka, Pahang* *Language Coordinator* *Roslina Abd Aziz*	TASK *Purpose is clear *Language use is appropriate but sometimes words std here not familiar with eg. Car boot sale *Time for preparation: rather short *Time for presentation: sufficient for both A & B *Length of test ok *Topics familiar, but quite difficult SETTING *Quite a problem, no special venue for spk test; tests in lecturers' office; noise, comfort, etc - Within campus: std tell each other topics; same set used in the day & evening -Between campuses, std tells each other too; test not conducted simultaneously	*Performance during the test is usually dep on the course- Dip Planting, Wood Tech std v weak - Dip Acc, Business quite good * Weak students: busy *translating* *Good std perform as expected; use of words/ expressions learnt	* Problem getting 2nd examiner for test; one examiner bias/more sympathetic to own students – problem with test at common time, time tabling * Rating criteria NOT described to students in detail *Rating criteria/ descriptors: too brief, not detailed like MUET's * Training: -MUET experienced lecturers well-versed - Last semester w'shop on the spk test; hands-on experience grading speaking test -New lecturers req briefing - Examiners compare & moderate marks	*Many lecturers prepare students for test; students unable to communicate well after that. - Syllabus related; previous syllabus allowed for more practice, presentations; students have more confidence to speak *Test impact on students is negative.		Fair test? *Not really testing the students' true ability to communicate *Not v fair to test them in a short time and 15 marks in that time *Students very focused on the test, how to score, etc., not thinking about whether they can really speak English or not * Should have a few tests, not just one final like this

Feb-11 UiTM Segamat, Johor *Lecturer/ Examiner* *Lizana Abdullah Md Ikram*	TASK *Content/topic familiar but they were not able to do well *Language ability/Content knowledge: - Syllabus related; it is MUET driven, not at improving proficiency, performance etc. - From basic Eng in BEL100, to 200, then jump to 250; big difference in topic difficulty~ here v challenging ones compare to earlier courses -Some topics more diff than others; also within topic **SETTING Test conduct** *Because it's tailored after MUET, it's appropriate~ many students NOT able to cope with time constraint: 2 minutes for preparation **Uniformity** Follow guidelines from SA **Security** Students discuss topics some times, but they don't do any better...	***Processing** - Not just language use; thinking skills, create presentation from question statements, & discussion amongst friends/peers -*Translating* from start, Malay to English/ thinking, speaking *Our students listen a lot, they write, but speak very little; limited language use	***Rating process** - 3 criteria but some do impression; some do benchmarking; *Subjective & varied.. *Descriptors: vague words eg Modest, good, superior etc. ***Training** Workshop on speaking ~ new lecturers ex-teachers, MUET examiners; difficulty rating our students	***Impact** *Preparing students too much for the test~ this sem. list of 22 topics frm SA, used & brainstormed in class *Students still couldn't spk well enough in the test	*Speaking test scores compared to scores in other components; no correlation *Adjust spk scores in borderline cases, esp. failing std	*Fair? -Other ways of testing spk; a few and a variety of assessments; ~ Time to identify weakness & work on improvements ***Other factors** - Knowledge based test so topics should be of students' interest - Group make up/familiarity with other speakers

Feb-26 *UiTM Shah Alam* *Lecturer/ Examiner & Resource Person* *BEL250* *Foziah A* *Raishah AH*	**TASK** *Purpose: dependant on the course; some need to do presentation (AP) more than others (PArts) in English *Preparation times not enough	***Processing** - Many students lack both content & language; have major problems expressing views, though task is very familiar	***Rating** -Rating criteria have descriptions; somewhat vague ~ satisfactory; modest.... - CA slight problem: enthusiasm? display confidence? gestures etc.	***Impact-** Preparing them from day one what's to happen in two month's time; **time constraint** in the semester -Brainstorming on topics, go through past yr papers, discuss strategies, etc.	*Speaking test marks: Cannot be adjusted; if student failing, look at final score & adjust Spk score if possible; consult class lecturer first	**Fair test?** - Yes, for the purpose of the test; for communication in other context, not certain/most can pass the test *Changes need to be made on syllabus, then the test; task A separated from task B, not together as it is now
	*Presentation time for group discussion not enough **SETTING** **Uniformity** *Some confusion about prep for group discussion ~ std do not read instructions carefully ~ some examiners allow discussion during prep + ~some lecturers group them before the exam * **Physical conditions**~ dep on faculty facilities; some better facilities/ classrrom conditions than others	- Thinking of points in L1 first, then L2 - Some actually do mapping/ branching of ideas - Most just write as much as they can; not much time to think carefully	- Mostly holistic marking, then only look at criteria	-Fulfil course objectives: No, because they forget everything after the test; following semester, same problems * We are zeroing in on the test....		

	DEMANDS *Topics familiar but students don't read enough, always a problem~ lack content knowledge *Language used in task appropriate					
Feb-26 University of Malaya, KL. Chief examiner MUET speaking test *Teoh Mei Lin*	TASK *Test has a very solid **purpose** ~ to speak/ conduct discussion in a given context ***Time** for preparation & presentation sufficient(2min) 3. **Moderators** agree on what represents a certain band (in training); agree on rating criteria & follow closely * **Response format** prob for private candidates; not enough prep on their own ***Demands** - Questions: within students range of experience; school & community related activities, projects, etc. - Topic: a few in a test; exam luck...Problems for those who don't read + more rural students	***Oral test limitations*** ~ unnatural speech ~ time constraint *These are candidates who need the **certification;** they are prepared, they are aware of what to do... **Content knowledge:** NO problem **Language**: MAJOR problem *Good speakers: Able to discuss, present views, interact; maximize preparation time * Weak speakers: - Trouble with words, utterances, in-complete sentences, ideas; prep time insufficient	***Scoring** - Terms used to describe criteria need to be fine tuned; 'very effective use' to 'limited use'... fairly good, quite good, etc. *Standardization during training- difficulty in determining bands 3-4; (5-6 very good, 1-2 very weak) ▪ TF they all score... **learnt** strategies ▪LA they lose most... problems with fluency, pronunciation, unclear speech ▪ CA follows; if LA low, CA probably low; learnt speech "I agree with you", "My instinct tells me..." * **2nd examiner** very important 3. **Uniformity** of marking: ▪Training helps ▪Who you select as	***In UM** test has had an impact* ~ min requirement to get in band 3; those with lower bands get in but have further language-skills courses... Otherwise unable to cope with other courses; -Onus is now on individual student, to improve their language skills or not		Fair test? *From a class-room perspective, **group interaction** approach is good; element of un-predictability, some natural discourse which you can test; in **individual presentation,** one can practise before hand.. *Nationwide perspective: too many factors to consider *Other factors: 1. Questions limited to a few issues; things they read in the papers...Should have topics/ issues more 'real' to students life 2. Examiners' own language ability/ competency

	-Nature of info: NOT factual/ content-based, more experience-based; some times abstract, eg. What is 'beauty' for you - Linguistics aspects: appropriate	* Private candidates: Format & procedure are unfamiliar; time spent reading & understanding question; prep time wasted *Very few: language is ok but not enough discourse; shyness, low self-esteem	examiners very important *MEC- exam board: Training- 2 weeks before exam, nation- wide, at 2 levels: national + state At present: reaching out to more teachers; not just for testing but also awareness on teaching, about rating criteria, etc.			3. Criterion CA needs revision- different CA for individual & group discussion
Feb-27 ***UiTM Shah Alam*** ***Academic Coordinator*** *Roseleena Md Noor*	**TASK** *Purpose clear: test student's ability in performing in a specific task, using approp language *Format: order works well; after individual pres, group collate info in a discussion to make decision; weak students benefit *Rating criteria made known to students by class lecturers *Time for prep & presentation is quite sufficient; test lasts about 20 minutes ***Administration**: - adopt MUET conditions - lecturers instructed to follow these conditions strictly	**Performance** *Weak stds memorize some words/ phrases ~ try v hard & get all wrong or don't speak; in discussion, try to interrupt *Good students say a lot, some even go out of topic and lose marks on TF *Students affected by both factors: - Language is a problem for most students - Content knowledge too; even with proficient std, don't read enough, lose out on content	**Scoring** *Band used adopted from MUET; 1 - 6 *Range is wide especially between faculties: most in range of modest user *Rating criteria comes with descriptors; clear & reliable in describing how a user, e.g. limited user, performs at a certain band, on a specific criterion, e.g. TF *2 examiners must be present at each venue; they compare & average marks *Problems when lecturers do not	**Impact** Most lecturers start preparing them early/ drilling - Speaking begins in semester 2, & reinforced in semester 3 ~ lots of practice ~ **acid test** on ability to interact with others in a group, not just on presentation skills	*Scores adjusted if discrepancies found *Total score reflects performance - if good in speaking, do well in other sections, & vice versa	*Fair test? - Yes, focus on ability to perform at the time, within a limited time; they all know what to expect, they had enough practice & content preparation *Other factors: - Topics must be carefully thought of; avoid bias & those that affect their emotions ~ too high flown to handle under testing conditions

	- Security: test conducted in a week; 2 sets of questions prescribed per day ~ random monitoring by coordinator, etc., walk into rooms at different faculties ***Demands:** -Length is appropriate -Nature of info: factual & current; familiar & related to life as students on campus -Language: avoid low frequency, difficult words; students range from very weak to excellent		follow guide-lines; very few cases *Briefings for all lecturers conducted from time to time; seminar on speaking assessment held last year *Core team on the speaking test: maintain same people to set, vet, etc for past 3 years			
Mar-02 *UiTM Shah Alam* *Deputy Dean, Language Centre* *Assoc Prof. Wan Latifah WA*	**TASK** *Purpose clear; what studens are to do specifically stated *Order is appropriate; in task B, std had formed opinion on topic to be able to contribute to the discussion *Time is sufficient, comparable to other situations when no real prep time is given; this is time to think & organize ideas **Setting:** *Tests conducted in regular classrooms;	**Performance** - Students' language skills & content knowledge vary dependant on b'ground & course ~ Good students have exposure, opportunity; weak students lack skills, content, language.... By test time, all students aware of what they have to do ~ to contribute to a	**Scoring** - Scoring guide is clear - Workshops at beginning of semester to keep lecturers posted on the various levels - Criteria: TF & LA clear; *CA not certain if everyone knows how to rate; need to be worked on, more explicit differences between individual & group discussion	**Impact** - Lecturers are given list of topics, fashioned after MUET topics, to help students prepare for test in terms of content knowledge - Syllabus driven, from proficiency-based; communication in social context to focus on communication in the work place, very	*Scores of different components put together; not enough analysis done - with more MUET scores, we are more conscious now to see if there are connections; in MUET, L & R scores high, W is lowest; Speaking test scores are improving; similar pattern with our students	*Fair test? - Yes because the objective is clear: to prepare them to speak in more academic and formal situations *Other factors: Feedback from industries is important for us, so we need to constantly look into other & newer methods of preparing them, & testing them

	dependant on faculty, some more conducive, but no audience, 4 candidates & 2 examiners only; environment familiar to the students *Uniformity of administration: - rating scales for both tasks A & B - 2 examiners - moderation at the end of term *Security: - 6 sets of situations/papers; different sets used from Monday – Friday; if suspect std knows the topic, examiner can change the set - One master copy sent out to branch campus; direct to campus director **Demands:** - Topics are familiar ~ students are exposed to them from readings & discussions in the classrooms	discussion, argue with supporting evidence, in an academic fashion, not emotions... most try hard to demonstrate this	- Moderation of class marks conducted every semester ~ most students score between 8-10 marks; too many scores that are much more or less is inconsistent	specific.... - Time constraint in the semester; teach speaking or teach for the test, lecturers do more than 2hrs allocated per week for speaking *In society: Most candidates able to get through; at least aware of what's required *In other class-rooms: They'll take time to overcome language difficulties, but also aware of what's needed there	now	
Mar 04 *UiTM Shah Alam* *Person-in-charge/ Leader of Speaking team, Language Centre*	TASK *Purpose is quite clear: They are to speak on a given topic, ensure the examiners and others understand their points, arguments, etc. *Two tasks	**Performance** *Students' performance have improved in **content knowledge**; a lot more exposure & reading on	**Scoring** *Examiners need to be more strict if it's based on MUET *Variation in rater assessments is big, especially	**Impact** * MUET had impact on the syllabus for English; revision to a more proficiency-based course in social &		**Fair test?** *Potential of a good test; not testing student's 'true' ability (no test can do that) Other factors: - Major concern

Padmini Menon	together show student's content knowledge, ability to communicate ideas sufficiently for others' to understand, & language accuracy & fluency *Rating is more straight forward, with MUET bands in view *3 criteria are necessary aspects of a speaking test *Limited time is necessary; within 4-5 minutes can gauge if students remain in band 3, or higher...	various topics in the classroom **language accuracy** still a problem for many * Students are getting close to what MUET expects them to do *They speak Malay all the time; language, format, skills for the test have to be taught...	across campuses ~ raters have own ideas of what bands 3 & 4 is.... *Training: - Lecturers shown how MUET marking is done - Seminars on marking writing & speaking - more standardized marking: rater A, cross the border, to rater Z, marking is the same...	academic contexts *In the English classroom: - Straight into format of test - Giving them intensive course on oral presentation & group discussion -Exposure to various topics in different forms: reading materials, internet sources, files on specific topics *Practical problems: time constraints of semester, students' lack of speaking in English, forces lecturers to prepare students for the test quickly - make them rehearse speech, words/phrases to use, etc.		on rater reliability across campuses - Other constraints beyond our control, e.g. too many holidays, shorten time for speaking especially (2 hrs a week); the students' unwillingness to participate in the target language
Mar 09 UiTM Shah Alam ***Lecturer/ Examiner*** *Norleza Manan*	**TASK** *Purpose is clear *Format – same every year; repeated *Time is very short, esp for preparation, so very important to	**Performance** - Group members are homogeneous; weak stds & good stds together - Weak stds do whatever they	**Scoring** - Two examiners; one is the class lecturer ~ rater bias; look for what's been taught to them + know students	**Impact** -Prepare them for test from BEL200, carried on to BEL250; same methods, focus on making full use of	**Speaking test scores** are not compared to scores from other components; they are reported separately, &	Fair test? - Yes, in terms of format & topics which students are very familiar with - Reflective of

	stress this; not to waste prep time thinking too much **Setting:** - Physical conditions: conducive esp conducted in class, during class time; students familiar with surroundings - Administration: Allow students to choose own groups; two examiners present -Security: Language Centre stipulates test sets for the day; some lecturers manipulate questions they use... **Demands:** *Topics familiar to them ~ from past years & discussions in class~ some are also not within their experience, e.g. community services, the National service	learnt or rehearsed; some strategies, words/phrases to use ~ struggle with language accuracy & fluency - Most students: lack content knowledge, even some proficient ones	very well - Examiners average scores & agree on final score; standardization - Raters: most are experienced; some in the MUET -Leniency in marking is dependant on which faculty the raters are from -Training: every now & then, the coordinator's initiative	preparation time for both tasks A & B - Stress the importance of how to present & what to say - Syllabus is MUET driven, & so are the tests	then added up for a grand total	their speaking ability ~ not really, esp when tested this way & with all other constraints of the semester Other factors: - Syllabus focus should not be primarily on MUET - A few speaking tests rather than a final one - Lecturers should pay more attention to what students do in the test -Not a good idea to test our own students; in MUET, candidates & examiners are not familiar

APPENDIX 3.6

GUIDELINES FOR THE INTERVIEW

A. Objectives of the interview

The main purpose of the interview is to gather first hand information from the candidates on the following aspects of the speaking test:

1. Strategies they use or employ in order to fulfil the test task (theory-based validity) in terms of preparation & presentation
2. b) Their thoughts/opinions/views on: test context (context validity); scoring method (scoring validity); the impact the test has made in the English class (consequential validity)
3. How these factors affected them during the test

This guideline was developed based on details outlined for 'Focus group' interviews (Krueger 2002), and is aimed at providing specific guidelines for the facilitator/ interviewer on the conduct and content of the interview.

B. Major considerations for the interview

• Participants

Students who have taken the speaking test: four candidates in a group

• Environment

The interview should take place in a room/classroom that is convenient for students to get to, with proper seating arrangements, and with minimum distraction (noise, people, etc). Other equipment/materials to be made available in the room:

- Video recording of student performance in the test
- Audio tape to record interview
- Guidelines for interviewer(s)

▪ Moderator/Facilitator

The interviewer(s) consist of: experienced English lecturers, the researcher and/or other interviewers skilful in managing group interviews and discussions

C. Conduct of the interview

The interview will be conducted as follows:

1. Introduction

- Welcome (Greeting, Introductions)
- Overview of the topic (Purpose of the interview in relation to the study)
- Ground rules (Discussion only about speaking test; try to be specific where possible; discussion is recorded; interviewer might take notes; feel free to express your thoughts/ideas;)

2. During the interview

- Ensure everyone is heard/ prevent one person from dominating;
- Keep the discussion on track
- Summarize points from time to time
- Be aware of 'group dynamics'
- Be alert to the time schedule
- Probe for elaboration such as

"What do you mean by that?"

"Would you explain further?"

"Would you give an example?"

- Record the discussion and/or take notes where necessary
- Control reactions to participants: verbal/ non-verbal, head nodding, short verbal responses (avoid "that's good", "excellent")

3. Conclusion

- Summarize with confirmation
- Review purpose and ask if anything has been missed
- Thanks and dismissal

D. Questions for the Interview

The questions are divided into three stages:

Stage I (Preparation)

1. How did you prepare for task A? What did you do?
2. How did you prepare for task B? What did you do?
3. More specifically, did any of the features below affect your **preparation** for task A? How did it/they affect your preparation?

Task setting: Purpose; Response format; Rating criteria; Time constraint; Physical conditions

Task demands: Nature of information; Test length; Topic/Content familiarity; Language range (words, structure, function)

Interlocutor (other speakers): Speech rate; Accent; Gender; Acquaintanceship

4. More specifically, did any of the features below affect your **preparation** for task B? How did it/they affect your preparation?

Task setting: Purpose; Response format; Rating criteria; Time constraint; Physical conditions

Task demands: Nature of information; Test length; Topic/Content familiarity; Language range (words, structure, function)

Interlocutor (other speakers): Speech rate; Accent; Gender; Acquaintanceship

5. Comparing your preparations for task A & task B, can you state any difference between them?

Stage II (Presentation)

1. What did you think about/do while you were presenting task A?
2. What did you think about/do while you were presenting task B?
3. More specifically, did any of the features below affect your **presentation** in task A? How did it/they affect your preparation?

Task setting: Purpose; Response format; Rating criteria; Time constraint; Physical conditions

Task demands: Nature of information; Test length; Topic/Content familiarity; Language range (words, structure, function)

Interlocutor (other speakers): Speech rate; Accent; Gender; Acquaintanceship

4. More specifically, did any of the features below affect your **presentation** in task B? How did it/they affect your preparation?

Task setting: Purpose; Response format; Rating criteria; Time constraint; Physical conditions

Task demands: Nature of information; Test length; Topic/Content familiarity; Language range (words, structure, function)

Interlocutor (other speakers): Speech rate; Accent; Gender; Acquaintanceship

5. Comparing your presentations for task A & task B, can you state any difference between them?

Stage III

1. What do you think of task A as a test of your speaking ability? Was it fair? Easy? Difficult? Support your answer with reasons or examples.
2. What do you think of task B as a test of your ability to speak in a group discussion? Was it fair? Easy? Difficult? Support your answer with reasons or examples.
3. Were there any differences between them? How are they different? Why?
4. Can you think of other factors that could have affected the way you performed in the test?
5. After a brief oral summary....

   Is that a fair summary of all that you've stated in this discussion?
6. After a review of the purpose of the study....

   Is there anything else that I have missed out that you would like to say/express?

Thank you very much for your cooperation and help in participating in this discussion.

Thank You.

UNIVERSITI TEKNOLOGI MARA
PUSAT BAHASA
JABATAN BAHASA INGGERIS

MAINSTREAM ENGLISH II BEL 250
SYLLABUS

PROGRAMME : DIPLOMA LEVEL ENGLISH

COURSE : MAINSTREAM ENGLISH II

CODE : BEL 250

CREDIT HOURS : 3

CONTACT HOURS : 6 HOURS PER WEEK

PRE-REQUISITE : MAINSTREAM ENGLISH I

COURSE DESCRIPTION

Mainstream English II (ME II) is the second of a two-semester English Language proficiency course designed specifically for diploma students of Universiti Teknologi Mara (UiTM). It builds on and furthe develops the major aspects of Reading, Writing, Listening, Speaking and Grammar. Students will have the opportunity to practise and integrate language skills in meaningful tasks relevant to an academic context.

COURSE OBJECTIVES

The course aims to raise students' level of proficiency in the English Language and to equip them with academic and critical thinking skills necessary to undertake tertiary studies. In addition, ME II prepares students to meet the requirements of the Malaysian University English Test.

COURSE CONTENT

The components of the course are Reading, Writing, Listening, Speaking and Grammar.

Reading — The Reading component develops students' ability to comprehend and interpret texts and non-linear texts through the following skills:

Review of
(a) Identifying topic sentences, main ideas and supporting details
(b) Distinguishing relevant ideas from irrelevant ones
(c) Paraphrasing
(d) Summarising (paragraphs of about 100 words)
(e) Drawing conclusions
(f) Transferring information from linear to non-linear texts (charts, tables, etc.) and vice-versa.

- Differentiating between fact and opinion
- Using the skills of intertextuality
- Making inferences
- Putting forward hypotheses
- Making predictions
- Extracting points to summarise linear/or non-linear text(s)
  [Students may be required to summarise only an aspect of the text(s)
- Analyzing and evaluating text
- Interpreting the writer's point of view

Intermediate and high-intermediate level texts will be used.

Writing

The Writing component prepares the students to write clear, well-organised academic prose employing the following skills:

- Review of (a) Developing thesis statements and topic sentences
  (b) Writing supporting details
  (c) Writing effective introductions, developmental paragraphs and conclusions
  (d) Defining concepts, explaining ideas, describing states and processes
  (e) Summarising paragraphs
- Organising and presenting information in logical order (e.g. chronological and spatial)
- Comparing and contrasting ideas, classifying information, and establishing cause and effect
- Drafting, editing, revising and rewriting to create the final draft
- Summarising information
- Responding critically and appropriately to information contained in texts and non-linear texts
- Documenting sources (bibliographic entries, references, etc.), acknowledging sources
- Identifying and overcoming common errors in writing

Students will be required to write expository and argumentative essays of 250 words and summaries of 100 words.

Listening

The Listening component trains students in the following skills:

- Review of (a) Listening for specific information
  (b) Listening for main ideas and specific information
  (c) Listening to paraphrases and summarise
  (d) Listening to take notes

- Drawing or interpreting information
- Predicting outcomes
- Drawing conclusions
- Recognising the functions of spoken language
- Analyzing and evaluating information

Various types of listening texts in social and academic situations will be used

Speaking

The Speaking component focuses on training the students to plan, organise and participate in effective individual presentations and group discussions It will involve t following

- Review of (a) Asking for and giving information
  (b) Asking for and giving opinion
  (c) Justifying opinion
  (d) Expressing agreement and disagreement

- Planning, preparing and delivering individual presentations
- Learning how to initiate, maintain and close group discussions
  --Presenting factual information
  --Making suggestions and recommendations
  --Stating and justifying opinions
  --Providing rationale for actions taken
  --Building a strong persuasive argument
  --Presenting alternate points of views
  --Asking and giving clarification
  --Accepting and rejecting ideas and proposals
  --Summarising and concluding

Grammar

The Grammar component helps students to gain accuracy in their language use The following items will be emphasized

- Review of (a) Passive Voice
  (b) Reported Speech
  (c) Clauses

- Using conditionals

The teaching of grammar will be integrated into all four language components

METHOD OF INSTRUCTION

Students will be taught through lectures, discussions, role-play, dictation, language games written tasks, reading tasks and other language enrichment activities. Audio and videotapes and computers will also be used

EVALUATION

	Component	Weightage (%)
Trial Exam	Reading Writing	} 0 %
Mid-semester Exam	Speaking	15 %
	Listening	15 %
Final Exam	Reading	45 %
	Writing	25 %
TOTAL		100 %

16 Jun 2000

REQUIRED TEXT

Pusat Bahasa, Universiti Teknologi MAra. 1999. Mainstream English II Kuala Lumpur: Longmans

JABATAN BAHASA INGGERIS
TEST SPECIFICATIONS
BEL 250 MAINSTREAM ENGLISH II

CONDITIONS	PAPER 2: SPEAKING (15%)
TEST OBJECTIVES	Using the right expressions in individual presentation and group discussion such as: { presenting factual information { making appropriate recommendations { stating and justifying opinions { expressing cause and effect relationships { persuading and drawing conclusions * presenting alternative points of views * accepting & rejecting ideas/proposals * asking & giving clarifications * summarizing & concluding Two (2) tasks: Formal situations based on the stimulus or topic given. TASK A - Individual presentation TASK B - Group discussion (involving three (3) to four (4) students at any one time Fewer or more than the number of students stipulated here, is not encouraged)
TEXT TYPE AND LENGTH OF TEXT	Social and academic contexts such as dialogues, discussions and presentations
PROPOSITIONAL FEATURES	Familiar/academic/social/general topics related to family, college and community issues
ORGANISATIONAL FEATURES	1. TASK A: Individual presentation – Two (2) minutes (e.g. giving suggestions/recommendations on possible activities that can be carried out by a school committee or a group of students for a particular event like donation drive or the English Language Week) 2. TASK B: Group discussion – 10 minutes (e.g. giving suggestions/opinions on a topic related to the event identified in TASK A, such as – "How can students be motivated to speak English in college?")
LANGUAGE LEVEL	Formal/Semi-formal
WEIGHTAGE OF MARKS	15 %
DURATION	TASK A: Preparation – Two (2) minutes Presentation – Two (2) minutes (students are required to talk one after another starting With Candidate A. Assessors should ensure that the stipulated time Limits are adhered to) TASK B: Preparation – Two (2) minutes Presentation – Ten (10) minutes (any one the candidates in the group may initiate the discussion. Assessors should ensure that the stipulated time is adhered to)
MARKING GUIDELINES	Please refer to the Score Sheet for Paper 2 (Speaking) BEL 200/250 (Note: The final mark are derived after averaging the evaluation of two assessors)

APPENDIX 3.7 (c)
Page A51

UNIVERSITI TEKNOLOGI MARA
FINAL EXAMINATION

COURSE	:	MAINSTREAM ENGLISH II PAPER 2 : SPEAKING
COURSE CODE	:	BEL250 (2)
DATE	:	
TIME	:	½ HOUR
SEMESTER	:	NOVEMBER 2004 – APRIL 2005

INSTRUCTIONS TO CANDIDATES

1. Candidates are required to complete **TWO** tasks i. e. **Task A** and **Task B**.
2. Each candidate is given two minutes to present **Task A**. (Each candidate is given two minutes to prepare **Task A**.)
3. Each group is given ten minutes to discuss **Task B**. (Each candidate is given two minutes to prepare **Task B**.)
4. Do not bring any material into the examination room unless permission is given by the invigilator.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO

This examination paper consists of 25 printed pages

SET 1

Candidate B

Instructions to candidates:

Task A: Individual presentation

- *Study the stimulus or topic given.*
- *You are given **two** minutes to prepare your responses.*
- *You are given **two** minutes to present.*
- *Listen to the others while they are making their presentations and take down notes for the group discussion in Task B.*

Task B: Group discussion

- *You are given **two** minutes to prepare points to support or oppose the other candidates' views.*
- *After you have listened to everyone, try to come to a decision as to which of the four suggestions is the best.*
- *Your group is given **ten** minutes for the discussion.*

Task A and Task B will be carried out consecutively.

Situation

UiTM is offering each college a RM50,000 grant to carry out a project to improve facilities on campus. Your college has set up a Planning Committee to decide on a suitable group project. As a member of the committee, give your suggestion.

Task A: You suggest that the grant be used to **improve hostel facilities on campus**. Give reasons to support your suggestion.

Task B: Discuss which of the following would be **the most useful** project.

& decide

The suggestions are:

(i) improving the transportation service on campus.
(ii) improving hostel facilities on campus.
(iii) setting up more computer centres on campus.
(iv) upgrading the sports centres on campus.

SET 1

Candidate D

Instructions to candidates:

Task A: Individual presentation

- *Study the stimulus or topic given.*
- *You are given **two** minutes to prepare your responses.*
- *You are given **two** minutes to present.*
- *Listen to the others while they are making their presentations and take down notes for the group discussion in Task B.*

Task B: Group discussion

- *You are given **two** minutes to prepare points to support or oppose the other candidates' views.*
- *After you have listened to everyone, try to come to a decision as to which of the four suggestions is the best.*
- *Your group is given **ten** minutes for the discussion.*

Task A and Task B will be carried out consecutively.

Situation

UiTM is offering each college a RM50,000 grant to carry out a project to improve facilities on campus. Your college has set up a Planning Committee to decide on a suitable group project. As a member of the committee, give your suggestion.

Task A: You suggest that the grant be used to **upgrade the sports centres on campus**. Give reasons to support your suggestion.

Task B: Discuss which of the following would be **the most useful** project.

The suggestions are:

(i) improving the transportation service on campus.
(ii) improving hostel facilities on campus.
(iii) setting up more computer centres on campus.
(iv) upgrading the sports centres on campus.

SET 4

Candidate A

Instructions to candidates:

Task A: Individual presentation

- *Study the stimulus or topic given.*
- *You are given **two** minutes to prepare your responses.*
- *You are given **two** minutes to present.*
- *Listen to the others while they are making their presentations and take down notes for the group discussion in Task B.*

Task B: Group discussion

- *You are given **two** minutes to prepare points to support or oppose the other candidates' views.*
- *After you have listened to everyone, try to come to a decision as to which of the four suggestions is the best.*
- *Your group is given **ten** minutes for the discussion.*

Task A and Task B will be carried out consecutively.

Situation

The recent increase in oil prices has resulted in many traders raising the prices of consumer goods without the government's approval. The Executive Committee of the Consumers' Association meets to discuss the most effective measure that can be taken to stop these traders from exploiting the situation. As a committee member, give your suggestion.

Task A: You propose that consumers be encouraged to **boycott traders who have increased the prices of goods**. Give reasons to support your suggestion.

Task B: In your group, discuss which of the following is **the most effective measure** to discourage traders from increasing the prices of goods.

The measures are:

(i) boycotting traders who have increased the prices of goods.
(ii) buying alternative products that are cheaper.
(iii) reporting traders who have unfairly raised prices to the relevant authorities.
(iv) writing in to newspapers to expose these traders publicly.

SET 5

Candidate A

Instructions to candidates:

Task A: Individual presentation

- *Study the stimulus or topic given.*
- *You are given **two** minutes to prepare your responses.*
- *You are given **two** minutes to present.*
- *Listen to the others while they are making their presentations and take down notes for the group discussion in Task B.*

Task B: Group discussion

- *You are given **two** minutes to prepare points to support or oppose the other candidates' views.*
- *After you have listened to everyone, try to come to a decision as to which of the four suggestions is the best.*
- *Your group is given **ten** minutes for the discussion.*

Task A and Task B will be carried out consecutively.

Situation

Reality television programmes like 'Akademi Fantasia' and 'Malaysian Idol' have gained great popularity in Malaysia recently. As part of a class project, you and your friends have been asked to identify one reason why such programmes are popular among Malaysian viewers. Give your opinion.

Task A: You feel that reality television programmes are popular among Malaysian viewers because they are **very interesting and enjoyable to watch**. Give reasons and examples to support your opinion.

Task B: In your group, discuss which of the following **contributes most** to the success of reality television programmes.

The reasons are:

(i) they are very interesting and enjoyable to watch.
(ii) they give viewers an opportunity to participate actively in the decision-making process.
(iii) the participants are ordinary people like the viewers and viewers can identify with them.
(iv) they give Malaysians the opportunity to showcase their talents and become famous.

UNIVERSITI TEKNOLOGI MARA
PUSAT BAHASA
JABATAN BAHASA INGGERIS

MAINSTREAM ENGLISH 1 & 2
BEL 200 & BEL 250
SCORESHEET FOR SPEAKING

	UiTM No.	Name	INDIVIDUAL TASK (A)			GROUP TASK (B)			Total (A+B)	-Final Score 30/2= 15%
			Task Fulfilment 6	Language 6	Comm. Ability 3	Task Fulfilment 6	Language 6	Comm. Ability 3		
A										
B										
C										
D										
A										
B										
C										
D										

Name of Examiner 1: _____

Signature : _____

Date : _____

Name of Examiner 2: _____

Signature : _____

Date : _____

INSTRUCTIONS ON THE SPEAKING ASSESSMENT AND SCORING PROCEDURE FOR BEL 200 AND BEL 250

1. Please study the assigned situations for each day and the criteria for the score guide given to you before the start of the exams.

2. Note that there are 2 tasks: Task A is Individual Presentation, Task B is group discussion

3. Each student should only be given **2 minutes** to prepare for Task A and Task B. They should write down notes in point form to save time

4. Ensure that the students concentrate on one Task at a time.

5. Both examiners must maintain a fair and objective marking system.

6. Study the criteria for scoring carefully to determine the good and weak students.

7. In instances where students do not understand the situations and need clarification, give them a general idea to get them started.

8. Examiners must bring sufficient blank paper into the exam room for students' preparation (**students are NOT allowed to bring in their own paper**).

9. There are **six sets** of the Speaking Paper. Please check that you have all six sets.

Order of the situations for each day

MONDAY	SETS 1 & 2
TUESDAY	SETS 3 & 4
WEDNESDAY	SETS 5 & 6
THURSDAY	SETS 1 & 4
FRIDAY	SETS 2 & 5
SATURDAY	SETS 3 & 6

Do not change this order under any circumstances.

Confidentiality of the exam:

Do not discuss any of the topics for the exams with the students.
Ensure that all students leave the premises as soon as they finish their exams.
Remember to follow the order of the given sets strictly. For example, students doing the exam on Monday will do only situations 1 and 2, Tuesday 3 & 4, and so on.
Please ensure that students do not leave with their rough notes when they leave the exam room. This will standardize the situations to be used all over UiTM campuses in Malaysia.

Recording of marks:

Please study the **revised Score Guides and Score Sheet (dated Feb 2000).**
Record your marks clearly on both examiners' score sheets and double-check each examiner's marks. **The final score is 15%.** Sign each other's score sheet. One examiner should assume the role of Chief Examiner in each team and ensure that the final score sheets are accurate.
Please record any problems and identify difficult or too easy situations for future reference.

UNIVERSITI TEKNOLOGI MARA
PUSAT BAHASA
JABATAN BAHASA INGGERIS

MAINSTREAM ENGLISH 1 & 2
SPEAKING SCORE GUIDE
(Individual Presentation)

SCORE / COMPONENT	6	5	4	3	2	1
TASK FULFILMENT (6 marks)	Fulfils task very competently	Fulfils task reasonably well	Fulfils task satisfactorily	Fulfils task modestly	Fulfils task in a limited way	Does not fulfil task
SCORE	6	5	4	3	2	1
LANGUAGE (6 marks)	Displays very confident control of language	Displays reasonably confident control of language	Displays satisfactory control of language	Displays modest control of language	Displays poor control of language	Displays very poor control of language
SCORE	3.0	2.5	2.0	1.5	1.0	0.5
COMMUNICATIVE ABILITY (3 marks)	Shows ability to communicate very competently	Shows ability to communicate competently	Shows ability to communicate satisfactorily	Shows ability to communicate modestly	Hardly shows ability to communicate	Does not show ability to communicate

MAINSTREAM ENGLISH 1 & 2: Task A (Individual Presentation) 6 + 6 + 3 = 15 marks

Feb 2000

UNIVERSITI TEKNOLOGI MARA
PUSAT BAHASA
JABATAN BAHASA INGGERIS

MAINSTREAM ENGLISH 1 & 2
SPEAKING SCORE GUIDE
(Group Presentation)

SCORE / COMPONENT	6	5	4	3	2	1
TASK FULFILMENT (6 marks)	Fulfils task very competently	Fulfils task reasonably well	Fulfils task satisfactorily	Fulfils task modestly	Fulfils task in a limited way	Does not fulfil task
SCORE	6	5	4	3	2	1
LANGUAGE (6 marks)	Displays very confident control of language	Displays reasonably confident control of language	Displays satisfactory control of language	Displays modest control of language	Displays poor control of language	Displays very poor control of language
SCORE	3.0	2.5	2.0	1.5	1.0	0.5
COMMUNICATIVE ABILITY (3 marks)	Shows ability to contribute to the discussion very efficiently	Shows ability to contribute to the discussion effectively	Shows ability to contribute to the discussion satisfactorily	Shows ability to contribute to the discussion fairly well	Shows limited ability to contribute to the discussion	Shows very limited ability to contribute to the discussion

MAINSTREAM ENGLISH 1 & 2: Task B (Group Presentation) 6 + 6 + 3 = 15 marks

Feb 2000

APPENDIX 3.8

TEST (DIALOGUE) SCRIPT Task B

Speaker	Dialogue	Action	Timing
A	Thanks for coming to the meeting. As you already know, visitors from Japan arrive next month. They would like to experience the Malaysian way of life as much as they can. I think we should think about places to visit, food, festivals or celebrations, and the education system in Malaysia. Do you agree?	'Faces' speaker B (both turned towards the test taker - so their faces can be seen)	Spoken at a speech rate of approximately 80 words per minute. Normal spacing between turns. No overlap. Screen freezes at end of question - face of questioner still on screen. **(BEEP)**
B	Yes, but maybe we should think about food first...Do you agree? If so, what kind of Malaysian food should we introduce the visitors to?	Same as above- Faces the test taker when asking the question at the end.	
TT		Test Taker (TT) speaks into microphone for 1 minute maximum.	**Gap on tape of exactly 60 seconds only. (BEEP)**
A	Great! We're all thinking along the same lines... let's think of some more ideas and discuss the details.	'Faces' speaker B (both turned towards the test taker - so their faces can be seen)	1 to 2 seconds movement on screen, then discussions begins again. Spoken at a speech rate of approximately 80 words per minute. Normal spacing between turns. No overlap. Screen freezes at end of second question - face of questioner still on screen. **(BEEP)**
B	OK. What about a talk on festivals or celebrations in Malaysia?	As above	
A	Hum... What do you think the talk should be about?	Turns to face the test taker when asking the questions	
TT		Test Taker (TT) speaks into microphone for 1 MINUTE maximum.	**Gap on tape of exactly 60 seconds only. (BEEP)**

B	That's a good idea! We could also take them on a visit to a school or a university.	faces speaker A (both turned towards the test taker - so their faces can be seen)	1 to 2 seconds movement on screen, then discussions begins again. Spoken at a speech rate of approximately 80 words per minute. Normal spacing between turns. No overlap. Screen freezes at end of second question - face of questioner still on screen. **(BEEP)**
A	But which one, a school or a university? OK, lets make a decision, which of the two should we go for, and what kind of things should they do there?	Same as above - BUT faces the test taker when asking the question at the end	
TT		Test Taker (TT) speaks into microphone for 1 MINUTE maximum.	**Gap on tape of exactly 60 seconds only. (BEEP)**
A	Ok...you spoke about places to visit, and we've just discussed things like food, a talk, and an educational visit. So, let's finally decide what to do. Can you summarize the ideas we've had and say what you think would be the best thing to do and why?	Faces the test taker (with B on the screen facing the test taker too)	1 to 2 seconds movement on screen, then speaker A starts. Spoken at a speech rate of approximately 80 words per minute. **(BEEP)**
TT		Test Taker (TT) speaks into microphone for 1 minute maximum	**Gap on tape of exactly 60 seconds only. (BEEP)**
A	Well, I think that was very good! Thank you for all your ideas and the decisions we've made today. Let's meet again next week.		

APPENDIX 3.9
CONTEXT VALIDITY QUESTIONNAIRE

Your opinions on the test	strongly disagree	disagree	no view	agree	strongly agree
1. Task A clearly states what I am required to do.	1	2	3	4	5
2. Task B clearly states what I am required to do.	1	2	3	4	5
3. Task A is a good test of my ability to communicate orally in an academic context.	1	2	3	4	5
4. Task A is a good test of my ability to speak English in social situations.	1	2	3	4	5
5. Task B is a good test of my ability to communicate orally in an academic context.	1	2	3	4	5
6. Task B is a good test of my ability to speak English in social situations.	1	2	3	4	5
7. Both task A and task B should have equal marks.	1	2	3	4	5
8. The criteria for scoring my performance (**T**ask fulfillment, **L**anguage use, **C**ommunicative ability) were made clear to me in the **test instructions**.	1	2	3	4	5
9. The order of the tasks, i.e. task A followed by task B is appropriate for the test.	1	2	3	4	5
10. In task A two minutes is sufficient time for a candidate to demonstrate his/her ability to present ideas.	1	2	3	4	5
11. The speaking time in task B is sufficient to present ideas	1	2	3	4	5
12. Having written and spoken instructions to prepare for tasks A and B is helpful.	1	2	3	4	5
13. The topic in tasks A and B is familiar to me.	1	2	3	4	5
14. The instructions for the tasks only contain words that are suitable for my level of language ability.	1	2	3	4	5
15. The instructions for the tasks use simple, easy to understand sentence structures.	1	2	3	4	5
16. The language functions I need to perform in the task (e.g. give opinion, give reasons, provide examples etc) are clear.	1	2	3	4	5
TASK A					
17. The **interlocutor** gave us extra help during the presentation.	1	2	3	4	5
18. I understood the interlocutor because s/he spoke at a pace that I could follow.	1	2	3	4	5
19. I find it difficult to understand the interlocutor because of his/her accent	1	2	3	4	5
20. I am happy with either a male or a female interlocutor.	1	2	3	4	5
21. I prefer the same gender interlocutor.	1	2	3	4	5

22. Knowing the interlocutor I am talking to makes me more comfortable.	1	2	3	4	5
TASK B	1	2	3	4	5
23. I was able to interact well with the other speakers because they spoke at *a speed that I could follow.*					
24. I am able to interact well with the other *speakers because* I understand the rules of turn taking.	1	2	3	4	5
25. I find it difficult to interact with speakers in task B because of their accent	1	2	3	4	5
26. I am happy working with a male or female speaker	1	2	3	4	5
27. I prefer to interact with the same gender speakers.	1	2	3	4	5
28. Knowing the speakers I am interacting with makes me feel more comfortable.	1	2	3	4	5
29. The following conditions in the test venue were OK:					
a) Lighting	1	2	3	4	5
b) Noise level	1	2	3	4	5
c) Room temperature	1	2	3	4	5
d) Seating arrangement	1	2	3	4	5
e) Conditions for disabled students	1	2	3	4	5
30. Candidates were not able to discuss exam questions with students who had already taken the test.	1	2	3	4	5
31. I prefer the computer version of the test to *the old test* format.	1	2	3	4	5

Interlocutor: Person you speak to/interact with in the test

Thank you for completing this part of the questionnaire.

APPENDIX 3.9
THEORY-BASED VALIDITY QUESTIONNAIRE

COMPUTER-BASED SPEAKING TEST: STUDENT QUESTIONNAIRE

Dear student:

We are conducting research on the university speaking test. This questionnaire aims to gather data as part of this research project. We want to get your feedback on the COMPUTER-BASED SPEAKING TEST which you have just taken.

This test was developed based on the UiTM or MUET speaking test. It is similar to those tests in the topic & tasks, i.e. it has both an individual presentation and a task in which you participate in an on-going dialogue.

The items in this questionnaire are divided into two aspects of the COMPUTER test. You will respond to questions related to the content of the test and what you did to perform the tasks.

The data collected from you will help us improve the existing test and provide valuable information for our on-going research. Therefore, your cooperation in the matter is greatly appreciated.

Thank you.

Sincerely,

Saidatul A Zainal abidin
Pusat Bahasa
UiTM Shah Alam

TASK A

What I thought of or did *before I started*	*strongly disagree*	*disagree*	*no view*	*agree*	*strongly agree*
1. I read the instructions very carefully to understand what was required.	1	2	3	4	5
2. I thought of the points I wanted to make.	1	2	3	4	5
3. I thought of how to satisfy the examiners.	1	2	3	4	5
4. I wrote down the points I wanted to make.	1	2	3	4	5
5. I thought of the words and expressions I needed to fulfil the task.	1	2	3	4	5
6. I thought of the structures I needed to fulfil the task.	1	2	3	4	5
7. I practised the speech in my mind.	1	2	3	4	5
8. I was familiar with the general topic of the task from previous readings and experience.	1	2	3	4	5
9. I knew enough specific information for the task from previous readings and experience.	1	2	3	4	5
10. The information in the instructions was necessary for me to complete the task.	1	2	3	4	5

What I thought of or did *during the planning time*	*strongly disagree*	*disagree*	*no view*	*agree*	*strongly agree*
1. I thought about the words and expressions I needed.	1	2	3	4	5
2. I thought about the grammar I needed.	1	2	3	4	5
3. I thought only in ENGLISH.	1	2	3	4	5
4. I thought only in MY OWN LANGUAGE.	1	2	3	4	5
5. I thought in both ENGLISH and MY OWN LANGUAGE.	1	2	3	4	5
6. I planned some ideas for each topic in my mind.	1	2	3	4	5
7. I was able to put my ideas or content in good order.	1	2	3	4	5

What I thought of or did *while I was speaking*	***strongly disagree***	***disagree***	***no view***	***agree***	***strongly agree***
1. When I spoke I checked:					
a) the appropriateness of the words that I used	1	2	3	4	5
b) my grammatical accuracy	1	2	3	4	5
c) the organization of my presentation	1	2	3	4	5
2. When I spoke I adjusted:					
a) the appropriateness of the words that I used	1	2	3	4	5
b) my grammatical accuracy	1	2	3	4	5
c) the organization of my presentation	1	2	3	4	5
3. I know I have to talk differently to a lecturer than to my friends in the class	1	2	3	4	5

Comments on the above items:

Thank you for completing this part of the questionnaire.

TASK B

What I thought of or did *before I started*	*strongly disagree*	*disagree*	*no view*	*agree*	*strongly agree*
1. I read the instructions very carefully to understand what was required.	1	2	3	4	5
2. I thought of the points I wanted to make.	1	2	3	4	5
3. I thought of how to satisfy the examiners.	1	2	3	4	5
4. I wrote down the points I wanted to make.	1	2	3	4	5
5. I thought of the words and expressions I needed to fulfil the task.	1	2	3	4	5
6. I thought of the structures I needed to fulfil the task.	1	2	3	4	5
7. My presentation in task A helped me complete this task.	1	2	3	4	5
8. The information in the instructions was necessary for me to complete the task.	1	2	3	4	5

What I thought of or did *during the planning time*	*strongly disagree*	*disagree*	*no view*	*agree*	*strongly agree*
1. I thought about the words and expressions I needed.	1	2	3	4	5
2. I thought about the grammar I needed.	1	2	3	4	5
3. I thought only in ENGLISH.	1	2	3	4	5
4. I thought only in MY OWN LANGUAGE.	1	2	3	4	5
5. I thought in both ENGLISH and MY OWN LANGUAGE.	1	2	3	4	5
6. I planned some ideas for each topic in my mind.	1	2	3	4	5
7. I was able to put my ideas or content in good order.	1	2	3	4	5

What I thought of or did *while I was speaking*	*strongly disagree*	*disagree*	*no view*	*agree*	*strongly agree*
1. When I spoke during the discussion, I checked:					
a) the appropriateness of the words that I used	1	2	3	4	5
b) my grammatical accuracy	1	2	3	4	5
2. When I spoke during the discussion, I adjusted:					
a) the appropriateness of the words that I used	1	2	3	4	5
b) my grammatical accuracy	1	2	3	4	5
c) the point(s) I wanted to make	1	2	3	4	5
3. When others are speaking, I checked the points they made	1	2	3	4	5
4. Based on what they said, I adjusted my next response	1	2	3	4	5
5. In the test, I had no problem doing the following:					
a) connecting what I said to what has just been said	1	2	3	4	5
b) concluding a group discussion	1	2	3	4	5
6. When talking in English to my classmates, I use different language than when I talk to my lecturers.	1	2	3	4	5

Comments on the above items:

Thank you for completing this part of the questionnaire.

REPORT ON WORKSHOP 'TESTING SPOKEN ENGLISH'
VENUE: UiTM SHAH ALAM
DATE : 28, 29 MAY 2005
FACILITATOR: PROFESSOR C J WEIR
ROEHAMPTON UNIVERSITY
LONDON, UK.

INTRODUCTION

Issues in testing spoken English have been in the testing literature for a long time, but validating the speaking test has not been common practice; the studies that have been conducted appear in different forms and are usually unsystematic (ref Weir 2004, Ch 2 for details on validity).

For this reason, a framework for validating the speaking test was developed which incorporates a socio-cognitive approach to validation, and five components of validity (Weir 2004/2005). According to the framework, the components we look at to ascertain if our test measures the construct it claims to measure are related to contextual features of the test task(s) (context validity), internal processing involved when candidates attempt the task (theory-based validity), and how the test is rated (scoring validity).

The following is an account of information gathered during the workshop on testing of speaking, including data obtained from participants who took part in rating exercises for the direct test and the computer-delivered test.

I. The following tables on *Aspects of Validity Components for the Construct 'Speaking' (Weir 2005)* **were presented as a basis for discussions on the topic**

Elements in the test task(s) which affect performance are:

Context Validity	
Setting: Task • Purpose • Response Format • Weighting • Known Criteria • Order of Items • Time Constraints	**Demands: Task** **Linguistic (Input & Output)** Mode Discourse mode Length Nature of information Topic familiarity Lexical range Structural range Functional range
Setting: Administration • Physical Conditions • Uniformity of Administration • Security	**Interlocutor** Speech rate Variety of accent Acquaintanceship Number Gender

All this in turn affects a candidate's cognitive & metacognitive processing, which then determines strategies he/she requires to perform the task through:

Theory-Based Validity		
INTERNAL PROCESSES • Conceptualiser • Pre verbal message • Linguistic formulator • Phonetic plan • Articulator • Overt speech • Audition • Speech comprehension	**MONITORING**	**EXECUTIVE RESOURCES** **Content knowledge** • Internal • External **Language knowledge** • Grammatical • Discoursal • Functional • Sociolinguistic

Aspects of how candidates are assessed include the rating criteria/scale, the rater, and the rating process

Scoring Validity

• Criteria/rating scale
• Raters
• Rating procedures
  ▪ Rater Selection
  ▪ Rater Training
  ▪ Standardisation/Accreditation
  ▪ Rating Decisions (inter-rater agreement)
  ▪ Consistency (intra-reliability)
  ▪ Moderation
• Grading and Awarding

II. Comments from participants based on the presentations and feedback received through questionnaires.

On context validity:

- Interlocutor factors such as speech rate, accent and acquaintanceship affect UiTM students' performance on the test
- Differences in language ability and degree of interaction for task B are apparent between faculties; students from Business school perform better than those in Applied Science where the examiner gives more help to encourage candidates to speak
- Importance of standardization in administration of test across faculties & branch campuses
- Students should be taught how to initiate, maintain and conclude a discussion

On scoring validity:

- Raters are interpreting each criterion differently; this is evident in the data:
  * on rating the **direct test** using the **existing scale** > the range in scores given were wide, showing a wide range of ability; we want this gap to be narrowed
  * on rating the **direct test** using the **new scale** > the range seems to have narrowed; an encouraging indication of the difference between this scale & the old scale

  (ref data outputs below)

Issues raised on rating:

- Existing scale not explicit enough
- Lack of standardization process among raters
- Separate scales for oral presentation & discussion
- Analytic scale with clear descriptors and reduced range

A. The following table is a summary of participants' feedback on the direct test and the computer test, based on Staff questionnaire responses

Test / Validity component	Direct test	Computer test
Context validity (Task A + B)		
Task setting		
Purpose	†	†
Response Format	†	†
Weighting	—	†
Known Criteria	—	†
Order of Items	†	†
Time Constraints	†	†
Task demands		
Linguistic (Input & Output)	†	†
Mode	±	†
Discourse mode	†	±
Length	±	†
Nature of information	†	†
Topic familiarity	†	†
Lexical range	†	†
Structural range	†	†
Functional range		†
Interlocutor		†
Speech rate	†	†
Variety of accent	±	†
Acquaintanceship		
Number	†	±
Gender	—	†
	±	†
Theory-based validity Conceptualizer/Planning (task A +B)		
Read task carefully	†	* NO ITEMS here for the computer test
Thought of points	†	
Wrote down points	†	
Satisfy examiners	±	
Thought of words & expressions	†	
Thought of structures		

Executive resource (task A +B)		
Internal knowledge	—	
External knowledge	†	
Checked linguistic knowledge	±	
Adjusted linguistic knowledge	†	
Monitoring (task A +B)		
Checked/Adjusted word use, grammar, organization of speech	±	
Checked effect of words on others; adjusted response based on other' points	† — ±	
Interactional functions/ management skills: initiating, keeping conversation going, connecting to what's been said, maintain turn taking	± ±	

Scoring validity		
Criteria/rating scale (appropriate, clear, sufficient)	±	†
Raters (trained, standardized, work well)	—	
Rating procedures		
▪ Rater Selection	—	
▪ Rater Training	—	
▪ Standardization Accreditation	—	
▪ Rating Decisions (inter-rater agreement)	±	Only (4) items here for scoring validity
▪ Consistency (intra-reliability)	±	
▪ Moderation		
Grading and awarding	±	
Statistical analyses on marks; 2 examiners present	—	
Committee checks test results	± —	

Key: † *Positive feature of test* — *Negative feature of test* ± *Uncertain about this feature of the test*

Note:

1. There appear to be some positive results for context and scoring validity for the computer test at this point, e.g. on scoring validity where upon actual rating of the tasks using the old and new criteria/rating scales, response were positive compared to rating the direct test using the old criteria/scale
2. Examiners were exposed to issues surrounding the direct test especially on rating, equivalence of input, standardization, co-construction of discourse.

B. The following summary is based on participants' **RATINGS** of the direct test & the computer test.

1. Rating the direct test using
    a) existing scale (RS1)
    b) new TOEFL scale
2. Rating the computer test using TOEFL scale
    a) by assigning overall score (RS2)
    b) by assigning score for each criterion (RS3)

DESCRIPTION OF THE OUTCOME OF THE RATING EXERCISE

The following conclusions have been made based on the rating exercise of this group of raters who consist of course lecturers, examiners, administrative staff and subject experts (ref Rating tables in Excel format)

1. On rating the direct test using the existing scale (RS1)
    - As a group, the overall range in scoring is large; very few raters with a narrow range in the scores for both tasks across four candidates
    - Some outliers (extreme scores; too low: 4, or too high: 12.5)
    - The objective is too make this gap narrower
    - Further discussion indicated raters interpreting a criterion (e.g. CA) differently

2. On rating the direct test using the new TOEFL scale

- A quick analysis of the data showed raters are rating much closer on the new scale in spite of absence of training and exposure to standardization tapes, than on a scale they had been using for some time

3. On rating the computer test using TOEFL scale
   a) by assigning overall score (RS2)
   b) by assigning score for each criterion (RS3)

   - Overall range is narrower than previously seen; very few inconsistency in scoring

4. On correlation : **all raters** rating **four candidates** on the **same test** using **different rating scales** > between rating the speaking test using the ITM scale and rating the speaking test using the TOEFL scale

   - Correlation indexes are low:

     For candidate 1 = 0.30
     2 = 0.54
     3 = 0.49
     4 = 0.26

   - No significant correlation or relationship in rating the candidates using the different scales

     - Rating scales are different in terms of clarity and specificity of descriptors

     - Range is wider when the test was rated using the ITM scale

     - Feedback from participants showed that ITM scale needs to be revised; each criterion needs to have clearer and more specific descriptions

     - As a group, raters seemed comfortable using the new TOEFL scale, in spite of lack of information and training

C. Feedback on the **COMPUTER TEST**

Features and potential benefits of the computer test

- Equivalence of input for all candidates in terms of instructions (speed, accent, linguistic & functional features are the same for all)
  *Addresses the issue of "co-construction of discourse" which affects performance & rating

- Equal time to prepare & respond
- Reliability of scoring (e.g. pass/fail cases can be easily identified)
- Interactional feature was evident between some candidates & input received; candidates' speech improved as they progressed through the task, especially in task B where there were four prompts for them to respond to ~ performance is incremental
- There is a permanent record; tests can be re-played & re-evaluated
- There is potential for standardization across the board, especially across campuses
- Task seems less stressful for candidates, especially with absence of examiner & other speakers
- Suitable for computer generation
- No need for interlocutor training

Potential problems of the computer test
- Effects on teaching; major change in teaching paradigm
- Infrastructure and cost
- Test security

CONCLUSION

1. **The direct test**
   Context validity:
   Data gathered from the questionnaires indicate that the areas of concern are weighting for each task, known criteria, nature of information, and interlocutor variables.

   Theory-based validity:
   Lecturers/examiners didn't agree that students were able to check and make adjustments to the structures they used in their speech, monitor their speech in relation to other speakers, and manage interactional functions during the group discussion task.

   Scoring validity:
   Response from questionnaires and discussions indicate that a major source of disagreement occurred for scoring in terms of clarity of criteria for rating, rater training & standardization, rating procedures such as consistency in rating and moderation, and grading and awarding procedures.

2. **The computer test**
   Context validity:
   Response here were favorable for all aspects of the test task

   Scoring validity:
   Response were positive for items related to criteria and scale for rating, and potential for intra as well as inter-rater reliability

Overall, there seemed to be some positive feedback for the computer-based test compared to the direct test; though more data is needed to further support these claims, for this group of participants who were exposed to both context as well scoring validity aspects of the tests, the impact on them were clear and positive.

There is yet a lot of work to be done and research to be conducted on the speaking test. Validation is the beginning of this process as we attempt to improvise our practices and the quality of our test. Validity is not a property of the test instrument itself, but it is what provides tests with quality as we concern ourselves with the soundness of the interpretations and uses we make of our test results. Using a socio-cognitive framework Weir 2004/2005) will enable the process to be more systematic and meaningful, especially since the relationship between the test task (context validity), the test taker (theory-based validity) and how we rate the performance (scoring validity) is crucial in determining construct validity. Further investigation is needed in areas of interaction in the speaking test (group work), on refining the computer test & gathering more data on it, and validation of the computer test.

Reported by:

PM Saidatul A Zainal abidin
Academy of Language Studies
UiTM Shah Alam

APPENDIX 3.11

INSTRUCTIONS (read and listen to)

This is **Task A** of the speaking test. In this task, you will play the part of a member of a club meeting at your university, and you are presenting your ideas. Here is the situation:

Some foreign exchange students are visiting your university next month. They will be spending a few months at your university hoping to learn a lot about Malaysia. Members of your club are holding a meeting to discuss interesting events or activities for the visitors.

For this task, you are required to talk about "Places to visit in Malaysia". In your presentation, you may include various ideas such as:

- The place to visit such as beach resorts, the city, or traditional villages
- Facilities available such as accommodation, restaurants/cafes, and shops
- Activities (indoor & outdoor) for adults and children
- Cost

Your score will be based on three criteria: **Delivery**, **Language Use** and **Topic Development.**
Delivery means how fluent you speech is and includes pronunciation and intonation
Language Use means how well you use grammar and vocabulary
Topic Development means how well you build and link your ideas and how well these relate to the task

You now have ONE minute to think about what you are going to say. **(Beep)**

(Beep) *You may now begin your presentation. You have TWO minutes to speak.* **(Beep)**

(Beep) *Stop talking now. Please look at Task B*

INSTRUCTIONS (read and listen to)

This is **Task B** of the speaking test. In this task, you will play the part of a member of a club meeting at your university. Here is the situation:

Some foreign exchange students are visiting your university next month. They will be spending a few months at your university hoping to learn a lot about Malaysia. Members of your club are holding a meeting to discuss interesting events or activities for the visitors.

In Task A, you talked about "Places to visit in Malaysia". In this task, you will discuss various aspects of the **'Malaysian way of life'** that you would like the visitors to experience. This includes topics such as:

- Food
- Festivals
- Celebrations
- Education system.

You now have 1 minute to think about these topics. **(Beep)**

(Beep) At the start of the next task, you will see and hear a discussion between two of the other members of the club. During this discussion you will be asked to:

- express opinions
- agree or disagree with statements
- help make decisions.
- summarise the discussion
- suggest what to do and why.

When one of the speakers asks you a question you will hear a (BEEP). Then you will have **a maximum of 1 MINUTE** to respond. A *second beep sound* (BEEP) will indicate when you should stop talking, and after that the discussion will continue.

You will be expected to respond four times throughout the discussion, each one with the same amount of time (1 Minute) for speaking.

(Beep)

APPENDIX 3.12

Computer Speaking Test Specifications (Context validity)

Note: This aspect of validity relates the test context, i.e. test setting, administrative factors and task demands.

CONTEXT VALIDITY	
Setting: Task • Purpose • Response Format • Weighting • Known Criteria • Order of Items • Time Constraints **Setting: Administration** • Physical Conditions • Uniformity of Administration • Security	**Demands: Task** **Linguistic (Input & Output)** Mode Discourse mode Length Nature of information Topic familiarity Lexical range Structural range Functional range **Interlocutor** Speech rate Variety of accent Acquaintanceship Number Gender

Specifications

SECTION	LEVEL: INTERMEDIATE/ADVANCE
General description of the computer test	• Use of computer interface/delivery; test conducted on the computer in computer labs with stand alone machines, one per candidate • Candidates responses are recorded in the installed MP3 recorder
Part 1: TASK A	Individual Presentation (Long turn /MONOLOGUE)
Part 2: TASK B	Interactive task (Short turns responding to prompts/ questions from interlocutor)

TASK TYPE	LEVEL: INTERMEDIATE/ADVANCE
PART 1: Task A	
Format	Long turn monologue
Number of candidates	One
Number of examiners	Examiner not present Interlocutor: Person giving instructions/information for the task
TASK DEMANDS	
Purpose	• To give the candidates the opportunity to speak individually • To test a student's ability to speak in the target language while demonstrating informational functions, based on the task assigned to him/her in task A
Response Format	Monologue presentation based on task provided on the computer screen
Known Criteria	Language use Delivery Topic development
Weighting	Both tasks are equally weighted
Time Constraints	1 minute preparation time; 2 minutes speaking time
Intended Operations	Informational
TEST DEMANDS	
Input	
No. of items	One
Channel	Written and spoken instructions
Discourse Mode	Informational
Text length *Test length*	A set of instructions followed by description of task, and suggested ideas/information for use in speech (Approximately ... words) Approximately 5 minutes (Inclusive preparation & speech times)
Nature of information	Non factual/ Not abstract Within students' experience, e.g. information related to college activities, the community, and so on
Topic familiarity	Parallel to topics in direct test
Lexical range	High frequency words; within students' lexical range at this level
Structural Range	Sentence structures familiar and within students' grasp at this level
Functional Range	Task requires candidate to elicit functions such as:

Time Constraints	1minute preparation time 4 minutes: 1 minute for each short turn response (4 times)
Intended Operations	Informational + interact ional
TEST DEMANDS	
Input	
No. of items	Four prompts/questions posed to student
Channel	Written and spoken instructions + dialogue between two speakers
Discourse Mode	Informational with some interaction
Text length *Test length*	A set of instructions followed by description of task, and suggested ideas/information for use in speech Dialogue between two speakers (Approximately ... words) Approximately 8 minutes (inclusive preparation & speech times)
Nature of information	Non factual/ Not abstract Within students' experience; linked to task A
Topic familiarity	Parallel to topics in direct test
Lexical range	High frequency words; within students' lexical range at this level
Structural Range	Sentence structures familiar and within students' grasp at this level
Functional Range	Task requires candidate to elicit functions such as in task A above, plus: Agreeing/Disagreeing; Suggesting; Expressing preference; Deciding; Summarizing
Content Knowledge	Within students' experience & scope of knowledge such as those relating to college life, community, and current events
Interlocutor	
Speech rate	100 - 130 words per minute
Variety of accent	Non-native speakers.
Acquaintanceship	Uncontrolled
No. of speakers	Two (interlocutors on screen)
Gender	Random
Expected Output	
Channel	Spoken
Text length	Not applicable
Lexical Range	At the level of intermediate to advanced English
Structural Range	At intermediate to advanced level; coherent and clear
Functional Range	Able to express opinions, justify the opinions, elaborate, describe ideas, agree/disagree, make suggestions, express preference, make decisions and summarize

Computer Speaking Test Specifications (Theory-based validity)

Note: This aspect of validity consists of the test taker and internal processing that the candidate goes through in attempting the test task.

THEORY-BASED VALIDITY		
INTERNAL PROCESSES • Conceptualiser • Pre verbal message • Linguistic formulator • Phonetic plan • Articulator • Overt speech • Audition • Speech comprehension	M O N I T O R I N G	EXECUTIVE RESOURCES **Content knowledge** • Internal • External **Language knowledge** • Grammatical • Discoursal • Functional • Sociolinguistic

Specifications

TEST TAKER DESCRIPTION	Test takers are candidates who are in semester 3 (Intermediate/Advance level) of the proficiency English course at the university
Physical: Gender	Male & female
Age	18 - 21
Accommodations	(Refer university documentation on accommodation for medical conditions, disabilities, students with special needs, etc.)
Psychological	Tasks types within candidates' experience and scope: oral presentations, question/answer task e.g. forum, group discussions/interactive tasks
Experiential: Content knowledge **Internal** **(Background knowledge)**	Topic/Information for test task/situation within candidate's scope and exposure: community-based, related to university life/activities, current events, etc.
External **(Information provided in task)**	Test instructions and task information/ situation appropriate and sufficient for candidates to be able to fulfill tasks adequately
Language knowledge **Grammatical**	Language output expected of candidates in terms of: - Understanding short utterances on a literal semantic level; includes phonology, stress, intonation, spoken vocabulary and spoken syntax

Computer Speaking Test Specifications (Scoring validity)

Note: This aspect of validity relates to how the test is marked and includes elements such as criteria/rating scale, rater, and rating conditions.

SCORING VALIDITY
• Criteria/rating scale • Raters • Rating procedures ▪ Rater Selection ▪ Rater Training ▪ Standardisation/Accreditation ▪ Rating Decisions (inter-rater agreement) ▪ Consistency (intra-reliability) ▪ Moderation • Grading and Awarding

Specifications

Scoring plan	Prior to marking all raters are to meet in a group to "norm" a sample audio recording of the computer test Each test recording must be rated according to the procedure listed below. Marks are to be recorded on a separate sheet. Marks are not to be shown to the second rater (or third, -when needed) until the rating process has been completed. All rating must be completed within the specified time limit
Criteria /rating scale	Every examiner will award a separate score for each criterion on the rating scale for each candidate. Analytic scale, criteria described in the realms of: Language use Delivery Topic development The criteria sheet contains a column on 'General description' + 3 criteria and each criterion has a scale from 0 - 4.
Rater Criteria	Qualifications: will have completed in-house rater-training All examiners will be required to participate in a norming session prior to the marking of the tests
Rater training &Standardisation	Assessment Supervisors to carry out face to face or online rater training at the beginning of each academic year. A training package will be developed

Rating Conditions	Assessors rate each recording independently and must not refer to previous raters' scores
Rating procedures	First marker uses analytic writing bands; second marker rates holistically corresponding to the writing band scale for that level. If the first marker and second are no more than 5 percentage points apart, split the difference between the marks. So, if marker A assigns 68% and marker B awards 71%, assign a final mark of 69.5 %. If the difference between the marks is greater than 5 percentage points, second marker fills out criterion form in detail. First and second marker discuss the presentation in order to reach a consensus. If not possible, refer to a third marker, who should perform a detailed final mark (i.e. not global).
Grading & Awarding	On completion of the grading process faculty are required to enter the grades into an Assessment Database. The grades will then be reviewed and finalized by Assessment Supervisors On completion of the review process, faculty will be notified that they can enter the grades into examination system. Faculty is required to enter the grades within the specified time limit.
Moderation	A minimum of 10% of all presentations will be marked by a moderator from another campus. A random sample of individual rater's scores will also be moderated after each test administration.
Statistical analysis	The reliability of the rating procedure will be estimated using both correlations and assessor agreement statistics (percentage agreement)